Statistical Mechanics and Local Dynamics of Denaturated Proteins

by

Ayşe Meriç Ovacık

A Thesis Submitted to the
Graduate School of Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Computational Science and Engineering

Koç University

July 2005

Koç University
Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Ayşe Meriç Ovacık

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

____

Burak Erman, Ph. D. (Advisor)

____

Atilla Gürsoy, Ph. D.(Co-Advisor)

____

Çagatay Başdoğan, Ph. D.

____

Metin Türkay, Ph. D.

____

Özlem Keskin, Ph. D.

Date: ____________________________

To my grandmothers, Süheyla and Fevziye

ABSTRACT

Bioinformatics is a fast growing research area and describes any use of computers to handle biological information. One of the major research efforts is structure prediction of proteins. The torsion angles (phi-psi) of proteins are considered as the degrees of freedom of a protein because of their control of the proteins' three dimensional structures. In this thesis, we used rotational isomeric state model in order to calculate statistical averages and correlations for torsional angles of denaturated proteins. For this purpose, we grouped each consecutive three residues (triplets) starting from first and used molecular dynamics simulations on triplets. Afterwards, we constructed energy maps for the phi-psi angles of the central residue of each triplet considered. Results showed that triplets have intrinsic propensities for some conformational preferences which favor the choice of the native state torsional angles and they are context dependent, determined by the amino acid sequence of the protein. Furthermore, we improved the stochastic weights with the aim of introducing the long range effects in two different approaches: Monte Carlo method and genetic algorithm method. Besides, we calculated heat capacity as function of temperatures by statistical mechanics for three different sets of stochastic weights which are obtained from molecular dynamics, Monte Carlo method and genetic algorithm method. Additionally, we proposed a dynamic rotational isomeric state model analogous to rotational isomeric state model and calculated transition probabilities from one state to another. The states are chosen as alpha-helix, beta-sheets, turns and all other states. Results support the idea that during folding, secondary structures forms sequentially.

ACKNOWLEDGEMENTS

I am deeply indebted to my advisor, Professor Burak Erman, not only for his constant support and great guidance but also for his patience. I am also grateful to my co-advisor, Professor Atilla Gürsoy. It is questionable how good codes I am writing, but it is not questionable that I have been thought how to write good codes by him. Without their help, this work would never be possible.

I would like to thank the members of my committee who accept to attend my defense: Profesor Çağatay Başdoğan, Professor Metin Turkay and Professor Ozlem Keskin. Their valuable reading, advice and patience are appreciated.

I would also like to thank to my CMSE friends, who are suffered from bioinformatics together with me. It would be a loss not mentioning their name here, Bora Erdemli, Sefa Nur Demirağ, Uğur Güner, Erk Subaşı and Volkan Kurt. And special thanks to Ahmet Bakan not only for his help but also for his friendship.

There are several people who would not understand the concepts in this thesis at all but without their being, this thesis would never be completed. I would like to thank Ayçıl Çeşmelioğlu for being everything to me, İrem Hisli for being my best friend and to Kaya family for being another family to me.

I would like to thank to Professor Pınar Çalık for making me believe in love, life and science.

Finally, I would like to thank my parents and sister. It is not possible to express my gratefulness in any words, but I can tell how much I fell lucky to be a part of Ovacik family. Special thanks to Nergiz Ovacik, for being the coolest and most caring mother in the world.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

MD	Molecular Dynamics
NMR	Nuclear magnetic resonance
RIS	Rotational Isomeric State
DRIS	Dynamics Rotational Isomeric State
PDB	Protein Data Bank
CI2	Chymotrypsin Inhibitor 2
X	First residue of the triplet
Y	Middle residue of the triplet
Z	Z third residue of the triplet
φ	the angle of right-handed rotation around N-C$^{\alpha}$
ψ	psi angle is the angle of right-handed rotation around CA- C$^{\alpha}$
ω	the angle of right-handed rotation about C-N bond
$E_{(\phi)}$	Total configurational energy
$u_{\zeta\eta;i}$	statistical weight for bonds i-1 and i in the state ζη
U_i .	Statistical weight matrix for bond i
$\Omega_{(\phi)}$	statistical weight of a configuration of the chain
ANOVA	Analysis of Variance
Z	Partition Function
E	microscopic energy
U	Internal Energy
k	Boltzmann constant
C_p	Heat Capacity
F	Generator Matrix
r_{kl}	Distance between bond r and k (Angstrom)
S^2	Radius of gyration (Anstrom^2)
n_p	number of virtual bond
l_p	Length of the virtual length

Chapter 1

INTRODUCTION

Proteins are polymer chains that consist of amino acid repeat units. They fold into unique 3-dimensional structures. The shape into which a protein naturally folds is known as its native state and is determined by its sequence of amino acids that are joined by peptide bonds.

A peptide bond is a chemical bond formed between two amino acid residues (There are twenty amino acids encoded by the standard genetic code) when the carboxyl group of one residue reacts with the amino group of the other residue, releasing a molecule of water. The chemical reaction of this process is as follows:

$$NH_3^+CR_1HCOO^- + NH_3^+CR_2HCOO^- \rightarrow NH_3^+CR_1HCOONH_3CR_2HCOO^- + H_2O$$

The chain obtained by this condensation reaction is shown in Figure 1.1

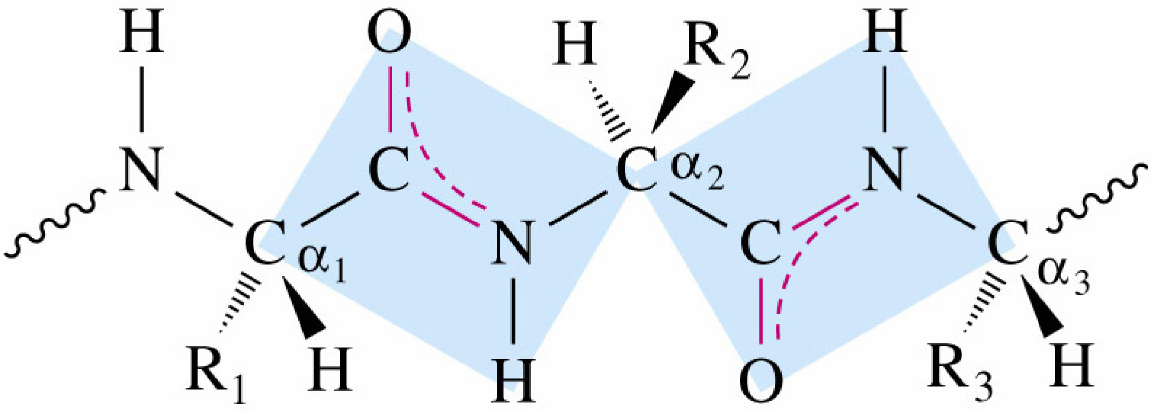

Figure1.1. The planes and center of rotation around a peptide bond adopted from [1]

The C-N bond has a partial double bond character (the nitrogen atom attaining a partial positive charge and the oxygen atom a partial negative charge) and rotation is not possible around this bond. The whole arrangement of C, O, N and H atoms is planar. The angle ω of the bond between C^{α} and N (generally close to 180 deg), the dihedral angles φ (the bond between N and C^{α}) and ψ (the bond between C^{α} and C) can have a certain range of possible values [2].

A torsion angle between 2 atoms j and k needs to consider 4 atoms i, j, k, l so that atoms i, j, k define one plane and atoms j, k, l define another plane; bringing the planes into coincidence gives the torsion angle along the bond where the planes intersect (the definition of torsion angle ψ (C_{α}-C) bond is given in Figure 1.2). In terms of defining a torsion angle for the main chain atoms of a protein, for the atoms N_1-C_{α}-C-N_2 the torsion angle is positive when the N_1-atom is made to coincide with the N_2-atom by a clockwise rotation when looking along the C_{α}-C bond. Consequently, the coordinates of the main chain atoms are used to calculate the dihedral angles φ and ψ. φ will be missing for the first residue in each chain and φ will be missing for the last residue in each chain.

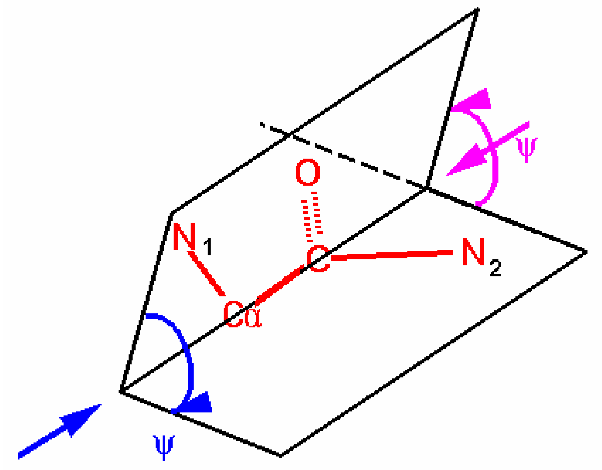

Figure 1.2. Torsion angle definition adopted from [1]

These angles are considered as the degrees of freedom of a protein because of their control of the proteins' three dimensional structures. The conformations of proteins can be determined from backbone configurations by specifying of proteins' φ and ψ angles.

One end of every polypeptide chain, called the amino terminal or N-terminal, has a free amino group. The other end, with its free carboxyl group, is called the carboxyl terminal or C-terminal.

Proteins are not only the building blocks of the cells but they also execute nearly all cell functions. The multiplicity of functions performed by proteins arises from a huge number of different three dimensional shapes which they adopt.

The particular amino-acid sequence of a protein causes it to fold into its native conformation. The folding process depends on the protein sequence as well as on the characteristics of their surrounding solution ([3], [4]) and temperature [5].

It appears that in transition to the native state, a given amino acid sequence always takes roughly the same route and proceeds through roughly the same number of fundamental intermediates [6]. At the coarsest level, folding first involves the establishment of secondary structure, particularly alpha helices and beta sheets, turns and only afterwards the tertiary structure [7]. In Figure 1.3., you can see the secondary structures and their chemical structures. The hydrogen bond formation between certain oxygen and hydrogen atoms identifies the secondary structures alpha-helix and beta sheet.

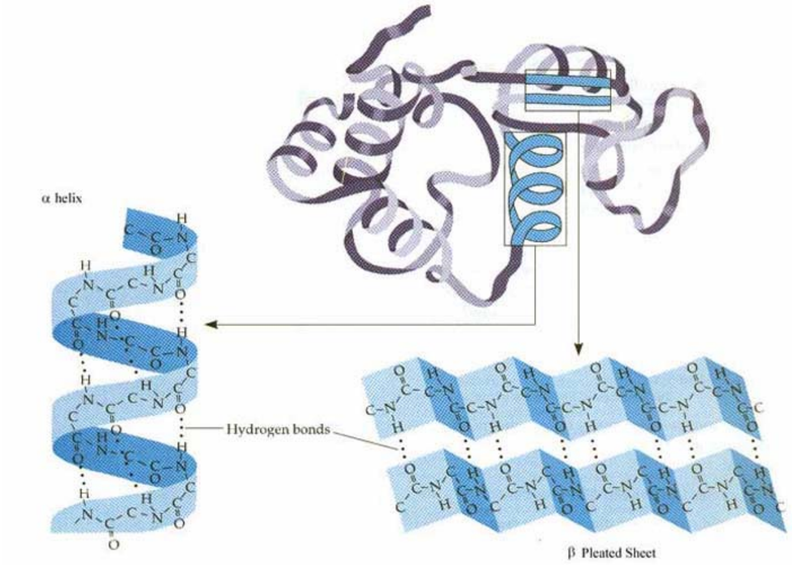

Figure1.3. Secondary structures in protein adopted from [1]

Ramachandran map specifies some of the secondary structure in terms of rotational angles. In Figure 1.4., we can see some of the specified secondary structures.

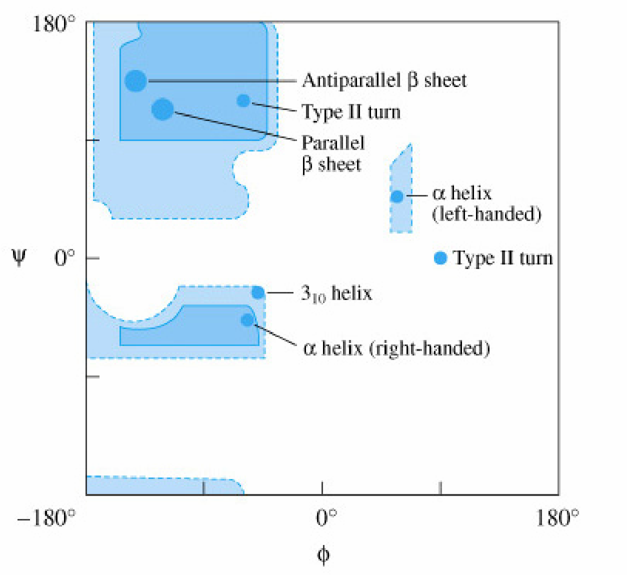

Figure1.4 Secondary structures on Ramachandran map adopted from [1]

Molecular dynamics (MD) simulation numerically solves Newton's equations of motion on an atomistic or similar model of a molecular system to obtain information about its time-dependent properties [8]. Beginning in theoretical physics, the method of MD gained popularity in material science and since the 1970s also in biochemistry and biophysics. It serves as an important tool in protein structure determination and refinement. The interaction between the objects is either described by a force field (classical MD), a quantum chemical model, or both [9].

Contribution

Several studies were performed about different levels of correlations among φ-ψ angles [2, 10]. NMR results also give clues about proteins' conformation preferences [11, 12]. In the present study, we show the context dependency of torsional angles by using MD. We used Chymotrypsin Inhibitor 2 (CI2) because of availability of the experimental studies. Moreover, we adopt rotational isomeric state (RIS) model to proteins by representing torsional angles as discrete states and try to predict heat capacity of CI2 which is one of the important thermodynamic properties via statistical mechanics. Additionally, we analyzed local dynamics of denaturated CI2 by the help of dynamic rotational isomeric state (DRIS) model.

Outline

Chapter 2 summarizes the previous studies that are related to our work. Chapter 3 elaborates the methods and main ideas used in this thesis. The model mainly depends on representing torsion angles as discrete states and the potentials are calculated as probabilities derived from the molecular dynamics. Chapter 4 illustrates the analysis of the probability calculations. Application of the statistical mechanics and heat capacity calculations are analyzed in Chapter 5. Last analysis is dynamics of denaturated CI2 and discussed in chapter 6. Chapter 7 finalizes the thesis by concluding the entire study.

Chapter 2

RELATED WORK

In this chapter we summarize the previous works about effect of neighbor residues to present residues' conformational preferences, theoretical and experimental studies on thermodynamic properties of CI2, and dynamics of proteins.

2.1 Short-range Interactions between Residues and Torsional Angle Preferences

Ramachandran put forward the correlations between φ and ψ angles of a single residue. His point of view included exclusion of steric overlaps which hold both denaturated and native proteins [2]. In this study, we extend the analysis of phi-psi angles to a sequence of triplets and include torsional energies into our calculations in addition to Ramachandran's steric maps. For this purpose, we use molecular dynamics simulations on triplets and construct energy maps for the phi-psi angles of the central residue of each triplet considered.

Interactions among neighbor residues are analogous to short-range interactions along the primary sequence. These interactions are not sufficient to give information about the tertiary structure of proteins as was discussed by Bahar et al. [13]. Molecular dynamics calculations for the triplets in this study relate to short-range interactions.

Our efforts for obtaining a more detailed study of the phi-psi Ramachandran maps were in part motivated by the work of Karplus which showed that torsional angle distributions have more fine structure than is generally observed [14].

The first clear demonstration of neighboring residue effect was given by Penkett et al. [11]. They introduced coupling constants of peptides by NMR studies. Jha et al. studied structural propensities for alpha helices, beta sheets in a restricted coil library and they concluded these propensities are often strongly influenced by both the chemical nature and the conformation of neighboring residues, contrary to the Flory isolated residue hypothesis [15]. The physical cause of the neighboring residue effect was studied by Avbelj et al. [16].

Keskin et al. used RIS model to calculate correlations for torsional angles of CI2 with two approaches: The first approach is using knowledge-based pairwise dependent torsional energy maps from Protein Data Bank (PDB) [17] and second approach is collecting torsional angle data from random coil configurations. Knowledge-based potentials showed strong correlations between neighboring torsional angles and those correlations favored the selection of the native state torsional angles [18]. Our work puts forward another approach by using RIS model that is constructing statistical weight matrices from molecular dynamics.

2.2. Experimental and Theoretical Studies about Heat Capacity of CI2

Obtaining information about thermodynamic properties of proteins is quite important to understand their equilibrium. In the literature CI2 system has been extensively studied as a model protein system in computational studies because of its small size (64 residues) and abundance of experimental studies.

Jackson et al.. evaluated experimental results of the heat capacity change over the temperature and for denaturation of CI2 [19, 20]. In their further work, Jackson et al.. studied also enthalpy and entropy of unfolding of CI2 experimentally [21]

In addition to the experimental results, there are several computational studies on heat capacity and transition phase of CI2 including the thermodynamics of its folding.

Kaya and Chan computed the heat capacity change with temperature and determined the transition state for CI2 theoretically by using an analytical model which is a polymer-lattice model for short sequenced proteins [22].

Similar to Kaya and Chan, Micheletti et al. studied the determination of the folding transition temperature by monitoring the temperature at which heat capacity show a peak [23]. They used an equilibrium analysis of proteins with known native state structures and introduced the term “native state overlap” which characterizes advancement of folding to native state by monitoring the residues’ configurations. In our study, we calculated the corresponding changes in the radii of gyration in order to define the rapid volume transition using the method introduced by Flory for calculating the averages of physical properties over configuration spaces[24].

Moreover, Day and Dugget discussed in their study the sensitivity of the folding and unfolding transition state of CI2 to changes in temperature. They concluded that the structure of the transition state of CI2 does not change significantly under varying denaturation conditions [25].

Lazaridis and Karplus presented a method for estimating the heat capacity of proteins via molecular dynamics and gave results from which contributions heat capacity arises [26]. Hao and Scheraga proposed a statistical mechanical study in their work. They did not use specifically CI2 but used a lattice polymer method and analyzed statistical mechanics characteristics of the model with the help of Monte Carlo simulations [27]. In our study we also use statistical mechanics to compute heat capacity, however with RIS model. Related to Lazaridis and Karplus, Hao and Scheraga introduced the effect of the solution on heat capacity. They did not observe any transition state appearance while analyzing thermodynamics properties such as average enthalpy and average thermal energy.

2.3. Dynamics of Folding Process

Dynamics of protein folding is a wide research area. Eaton et al.. summarized in their experimental work the protein folding and unfolding on the previously inaccessible nanosecond-microsecond timescale. They concluded that the comprehension of the protein folding mechanism includes knowing the elementary structural processes in protein folding. Scientists are now able to measure rates on a wide range of timescales of interest by the help of available fast kinetic methods [28].

In addition, Fersht reviewed nucleation mechanisms in protein folding in his study. Fersht emphasized the Levinthial paradox which says that there are an enormous number of conformations open to the denaturated state of a protein and a search through these would take an eternity. Therefore, one can conclude that the folding process is hierarchic and there exist folding intermediates and transition states through folding [29].

Eaton et al. stated in their work that some of the folding intermediates fold formerly. Alpha helices fold fastest among other secondary structures [28]. Forcellino et al. proposed a simple model that characterizes the folding of small proteins. They questioned alpha helices and beta-sheet barriers affecting the protein folding problem [30].

It is important to understand the local dynamics of proteins at the residue level. There are few studies about local chain dynamics of proteins. Elizier et al. stated that under weakly folding conditions, the polypeptides fluctuate between unfolded states and local elements of structure that become extended and stabilized as the chain becomes more compact. These results provide a detailed model for molecular events that are likely to occur during folding of myglobin [31]. Moreover, Markwick et al. studied local structure and backbone dynamics and present strong correlations between NMR relaxation and local psi angle [32].

Skolnick and Kolinski studied 24-nearest-neighbor lattice model of proteins that includes both alpha and beta-carbon atoms. They examined number of distinct situations. They concluded that the universal conditions for the formation of a unique native conformation are tertiary interactions and the occurrence of relatively small intrinsic turn preferences that choose the native conformation from a numerous of packed states. They also concluded that the results are universal; they do not depend on lattice, protein model or Monte Carlo dynamics [33].

Ding *et al*. stated that denatured states indeed have strong local conformational bias toward native states while a random-coil power law scaling of protein sizes is preserved. [34].Furthermore, Kiefhaber *et al.* studied unfolded chain dynamics as a model for the earliest steps in folding mechanisms. They used triplet-triplet energy transfer to measure formation of interchain contacts in several proteins and they examined the effect of amino acid sequence on local chain dynamics by using host-guest peptides. They could determine the time constants for development of the earliest steps intrachain contacts during protein folding [35].

The RIS model is a powerful tool to investigate the configurational statistics of polymer chains. Bahar and Erman develop a description of local chain dynamics in terms of conformational transitions between isomeric states[36]. There are several articles that study the DRIS model and make a comparison with the previous and theoretical works [37], [38], [39]. However, all the previous works are for polymer chains. We apply the method in a different way for proteins. Previous works define rates from the energy differences in torsion energies while we use directly molecular dynamics simulations in order to compute the rates of transition between basins of Ramachandran maps. Via DRIS model we calculated transition probabilities for individual bonds.

Chapter 3

MATERIALS AND METHOD

3.1. Simulations

The Chymotrypsin Inhibitor 2 (2CI2) has 64 residues available in the Protein Data Bank (PDB) [17].In Figure 3.1 one can see the three dimensional structure of CI2. We grouped each consecutive three residues (triplets) starting from first and called them triplets by using VMD [40]. More explicitly, the organization of the triplets is as follows: first triplet contains 1st, 2nd and 3rd residues and second triplet contains 2nd, 3rd and 4th residues and etc. Therefore we have n-2 triplets where n is the number of residue. We denote the first, second (middle) and third residues as Let X, Y and Z, respectively.

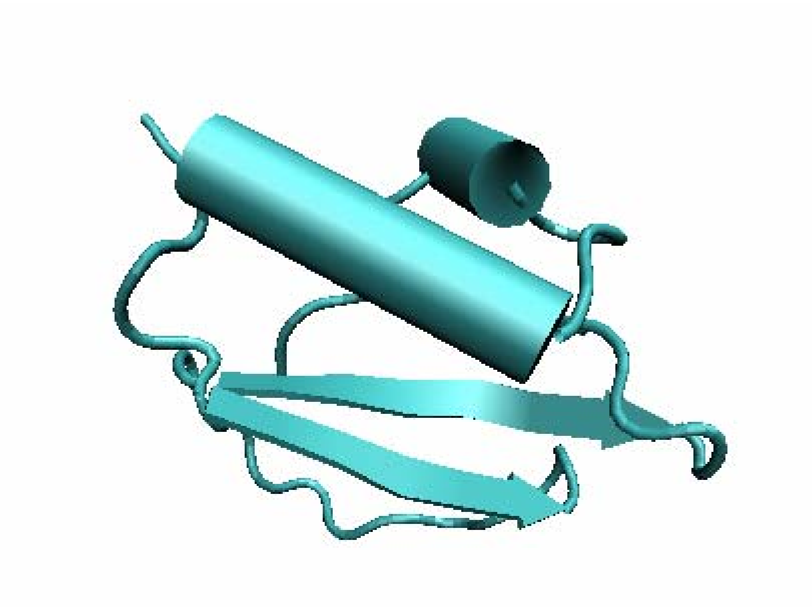

Figure 3.1. Tertiary structure of CI2 retrieved from VMD [40]

A triplet comprising the i-1st, i th and i+1st residues, has six torsion angles, ϕ_{i-1}, ψ_{i-1}, ϕ_i, ψ_i, ϕ_{i+1}, ψ_{i+1}. The torsion angle ϕ_{i-1} that indicates rotation about the N_{i-1}-C^{α}_{i-1} is undefined. Throughout the simulations the angles ϕ_{i-1} and ψ_{i+1} were fixed at 0^o. Similarly, the torsion angle ψ_{i+1} that indicates rotation about the C^{α}_{i+1}-C_{i+1} is undefined. It was also fixed at 0^o. The four internal angles ψ_{i-1}, ϕ_i, ψ_i, ϕ_{i+1} define the conformation of the triplet.

Among these four internal angles, we focused on the pair-wise occurrence of the angles ψ_i- ϕ_i, and ψ_i- ϕ_{i+1}.

The first problem that we faced was the duration of the simulations. Four ns simulation of one of the triplets in a water box take approximately 27 hours which is very long computation time. Therefore we decided to simulate the triplets in vacuum but with a dielectric constant 20 which is used commonly for proteins [41, 42]. The simulations take c.a. 1.5 hour in vacuum for 5 ns.

The purpose of the simulations was to determine the preferences of the triplets for certain specific conformation via molecular dynamics. After few trials we saw that the initial configuration played crucial role for the conformational preference of triplets. Therefore, we defined eight initial states and set the initial conformations of each triplet to one of those states. The states are given in Table 3.1.

Table 3.1 the torsional angles defined as initial states

Initial State	ψ_i	ϕ_i	ψ_{i+1}
1	90^o	-90^o	90^o
2	90^o	-90^o	-90^o
3	90^o	90^o	90^o
4	90^o	90^o	-90^o
5	-90^o	90^o	-90^o
6	-90^o	90^o	90^o
7	-90^o	-90^o	90^o
8	-90^o	-90^o	-90^o

Another problem was raised when we performed trials for the simulations. For some initial states, the triplet was trapped around the initial states conformation due to hydrogen bonds and could not move freely. In Figure 3.2., nitrogen, oxygen and hydrogen has colors as blue, red and black, respectively. We see the hydrogen bond between the oxygen of the residue Z and the hydrogen bonded to nitrogen of the residue X blocks the triplet's motions.

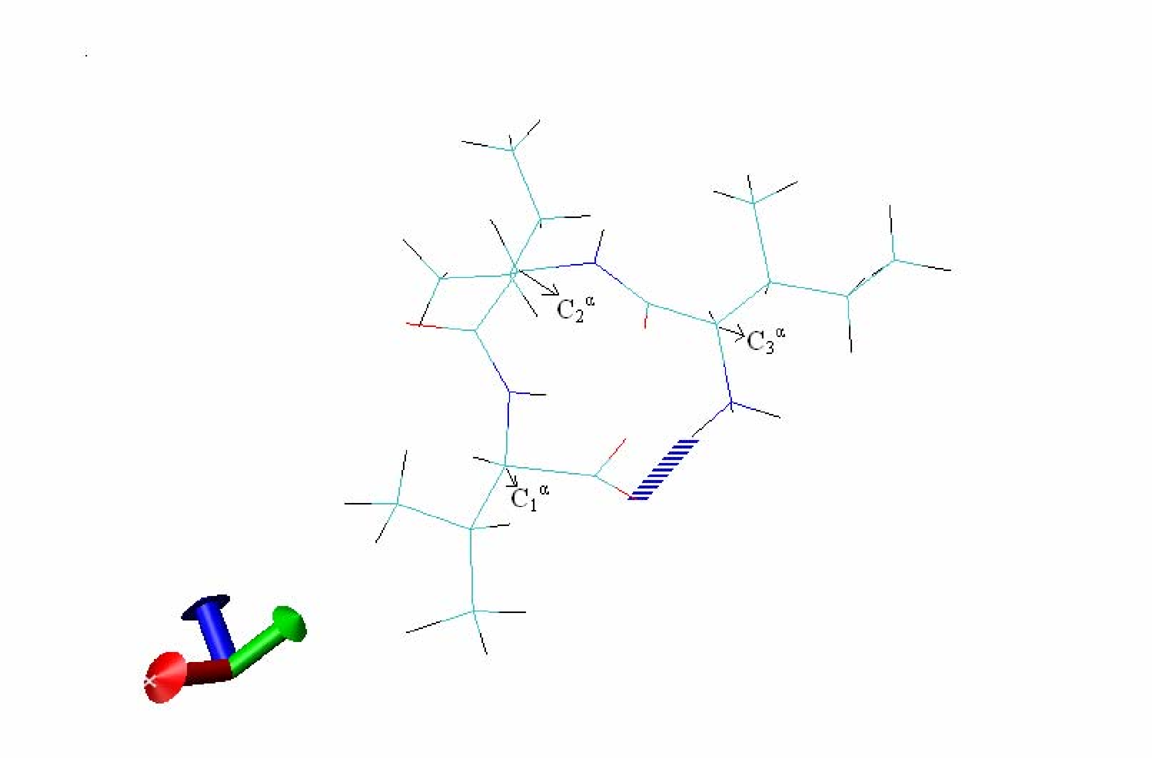

Figure3.2. Hydrogen bond formation due to initial condition of triplet

X-ray crystal diffraction usually cannot resolve the positions of hydrogen atoms or reliably distinguish nitrogen from oxygen and carbon, and over 80% of the three-dimensional macromolecular structure data in the PDB were obtained by X-ray crystallography [17]. Most of the MD packages use N-methylamine for N-terminal ending of protein and Acetyl group for C-terminal of protein [41-43]. We used the MD package NAMD [44] with CHARMM27 [42] parameter file where the default ending for

N-terminal is one hydrogen and for C-terminal one oxygen. CHARMM27 parameter file contain several C-terminal endings for any purpose of using [42], i.e.: Fanelli et al.. used methylated C-terminal in their study for docking purposes [45].

However, none of the C-terminals could make the H-bond avoid to be formed because of the present oxygen. In order to impede the H-bond formation between the oxygen of residue Z and the hydrogen of residue X, we converted carbonyl terminal to sp3 carbon (the carbon with 4 single bonds) with 3 hydrogen atoms as described in further detail in Appendix A.1.

Next step was to check whether the triplets' preferences favor certain basins in the Ramachandran plot or not, such as beta sheets or alpha helices. For this purpose, we examined the simulation results for specific triplets that found in beta-sheets and alpha helices in CI2.

Figure 3.3. shows the simulation results for the residues 34, glutamic acid; residue 35, alanine; and residue 36, lysine, that are found as alpha helix in CI2. Initial coordinates were set to beta-sheet conformation for that simulation which is one of the initial states we defined already. In Figure 3.3., the filled circles show the values of the torsion angles ϕ and ψ obtained at specific intervals during the simulations. We can see from the Figure 3.3 that the residue 35, which is the middle residue of the triplet, prefers not to be in the beta-sheet. On the contrary, it appears near the alpha helix conformation. The red button represents the area where the simulation ended.

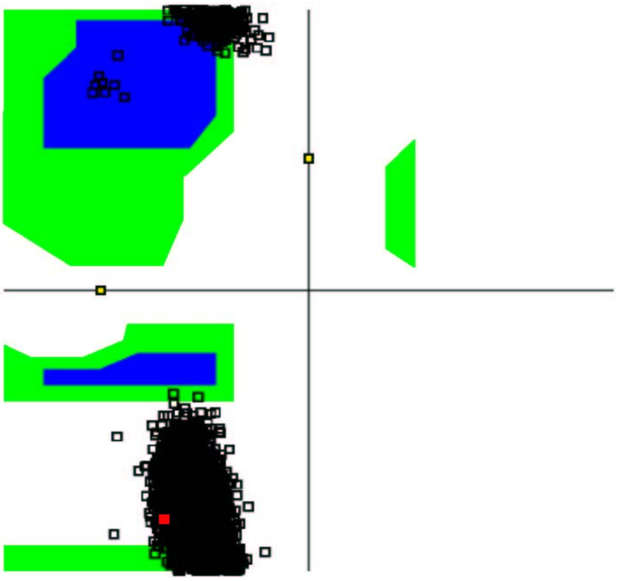

Figure3.3. The change of torsional angle configuration of residue 35 for the entire simulation

Similarly, we check the triplet that is in the beta-sheet conformation in CI2. The simulations results are presented in Figure 3.4. for the triplet that contains residues 48, 49 and 50. The residues are isoleucine, isoleucine and valine, respectively. Initial coordinates were set to the alpha helix conformation; however the residue prefers to be in the beta-sheet conformation.

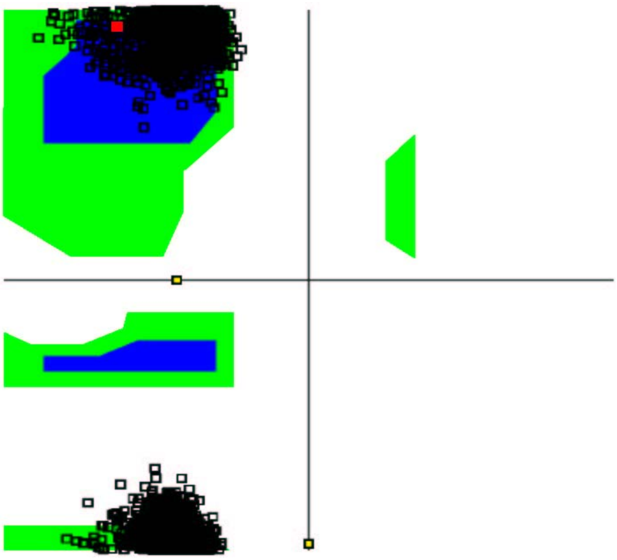

Figure3.4. The change of torsional angle configuration of residue 49 for the entire simulation

However, we do not expect all the triplets to prefer the right conformation for the CI2, since the triplets could be at any conformation in different proteins, a triplet could be in an alpha helix in one secondary structure whereas it could be in a beta-sheet in other protein.

In the simulations, the temperature was maintained at 310 K by means of Langevin dynamics using a coupling coefficient of 5/ps [42, 43]. Initial velocities were generated randomly at 310 K in accordance to the masses assigned to the atoms [42]. The time step was 1 fs and configurations were sampled at 1000 fs intervals. Minimization was done for 200-300 steps. Exclusion policy is defined as the value of 1-4, all 1-3 pairs will be excluded along with all pairs connected by a set of two bonds (i.e., if atom A is bonded to atom B, and atom B is bonded to atom C, and atom C is bonded to atom D, then the atom pair A-D would be excluded) [46].Cutoff distance and pairlist-distance were used as 12 Ǻ and 13.5 Ǻ respectively [46]. The simulations were carried out in a Linux-based cluster which each node has Intel Pentium 4 2.4 GHz processor.

Finally, we have 8 different 5ns simulations each of them corresponds the initial configurations for each triplet. 5000 data points were saved for each simulation and we have 40000 data points for each triplet.

3.3. Rotational Isomeric State Approximation

In the rotational isomeric state approximation, each bond is assumed to obtain several discrete rotational states. We use discrete state formalism for the torsion angles, where each torsion angle is divided into 30^o intervals. Therefore, we have 12 torsional states representing the torsion angles, for example: any angle between -180^o and -150^o is represented as state 1, while any angle between 180^o and 150^o is represented as state12. The basins indicated by blue are the beta and alpha regions in Figure 3.5.

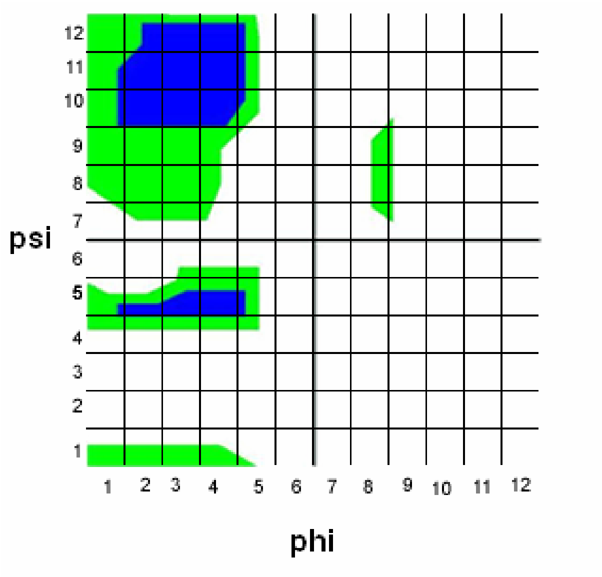

Figure3.5. Twelve states that represent each torsional angle

Description of the conformations of the molecules in terms of discrete rotational isomeric states is both convenient and well rationalized by physical circumstances [47].

If a given bond assumes a definite rotational state, the occurrence of either of its nearest neighbors along the chain in that definite rotational state is strongly encouraged. Therefore, the rotational potential for a given bond acquires dependence on the rotational states of its neighbors. The RIS approximation takes the place of all other characteristics of bond rotational potentials in its effect on the configurations of chain molecules.

3.4. Probability Level

By the RIS model we calculate the probability of each state for a bond. The method of calculation is outline in Appendix A.2. We clarify the concept with following example.

In Figure 3.6., we can see the probabilities of states for residue 24’s phi angle. 24th residue of CI2 is tryptophan and its native psi angle is 61.205 (9, in state representation.)

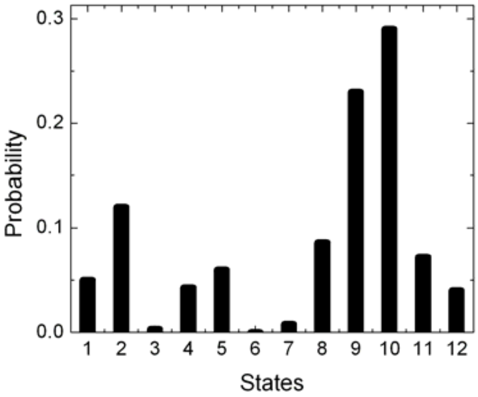

Figure 3.6. The probabilities for residue 24's psi angle

Then, we sort the states in descending order like in Figure 3.7

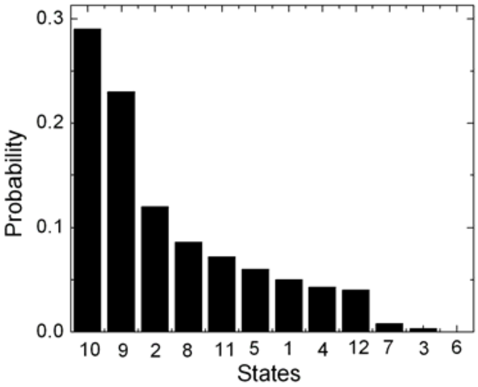

Figure 3.7. The probabilities for residue 24's psi angle in ascending order

After sorting the probabilities of states, we check the native state's order. The highest probability is state 10. If the native state of this angle would be 10, then probability level for this bond would be 12. However, the native state of this angle is 9, and state 9 has the second highest probability. Therefore, probability level for this angle is 11.

We define "the probability level" term in order to determine the achievement of RIS model for each bond easily.

3.5. Statistical Wight Matrices for Interdependent Bonds

Here, for demonstration; we have 6 bonds and use two states for each bond α and β. Assume that we have a configuration:

α α β α β β

The rotational potential affecting any given bond i, depends exclusively on configuration of i-1, i, i+1 bond. Interactions of longer range are ignored. Then, the total configurational energy can be expressed as a sum of energies for the first-neighbor pairs;

$$E_{(\phi)} = E_{\alpha} + E_{\alpha\alpha} + E_{\alpha\beta} + E_{\beta\alpha} + E_{\alpha\beta} + E_{\beta\beta} \tag{3.1}$$

In general, we may write the total configurational energy for n bonds as;

$$E_{(\phi)} = \sum_{i=1}^{n-1} E_i(\phi_{i-1}, \phi_i) = \sum_{i=1}^{n-1} E_{\zeta\eta;i} \tag{3.2}$$

where ζ denotes the state of bond i-1 and η that of bond i. $E_{\zeta\eta;i}$ is taken into consideration as the contribution to $E_{(\phi)}$ related to the assignment of bond i, to state η, bond i-1 being in state ζ.

The statistical weights are calculated from the conformational energies. The statistical weight for bonds *i-1* and *i* in the state *ζη* is given as:

$$u_{\zeta\eta;i} = \exp(-\frac{E_{\zeta\eta;i-1,i}}{RT}) \tag{3.3}$$

Statistical weights for all states of the bond pairs i-1 and i may be arranged in a matrix, called the statistical weight matrix, ***U_i***. The *ζη th* element $u_{\zeta\eta;i}$ of ***U_i*** represents the statistical weight when bond i is in state η while the bond i-1 is in state ζ. The statistical weight of a configuration of the chain as a whole is given by

$$\Omega_{(\phi)} = \prod_{i=1}^{n-1} u_{\zeta\eta;i} \tag{3.4}$$

Flory stated that the statistical weights could be chosen as the primary quantities for characterizing a configuration rather than energies [47]. Moreover, statistical weights must include neighbor dependence and yield correct statistical weight for any configuration of the bond.

The configuration partition function is given as;

$$Z = \sum_{(\phi)} \Omega_{(\phi)} = \sum_{(\phi)} \prod_{i=1}^{n-1} u_{\zeta\eta;i} \tag{3.5}$$

where the summations are taken over all configurations.

From the simulations we derived residue-specific conformational potentials as probabilities. The details of the calculations are given in the Appendix A2 and analysis of the calculations is given in Chapter 4.

Chapter 4

ANALYSIS OF PROBABILTY CALCULATIONS

4.1. Comparison of Probability Levels

The probabilities are calculated using the method described in the Appendix A.2. Since there are 12 states for each torsion angle, we can arrange these states from 1 to 12, in an increasing order with respect to their probabilities. We then identify for each bond for each bond, the probability level of the native state. For the ith bond, for example, if the highest probability is the same as in the native state, we identify the probability level of the ith bond as 12. We calculated the probability level of each bond in this manner. Results are shown in Figure4.1, where the triangles are the probability levels obtained from the triplets and the circles are obtained from the full sequence. The abscissa in the figure represents the bond indices of CI2. Since there are two rotatable bonds for each residue, the number of the bond indices is twice that of the residues number. The ordinate corresponds to the probability level for that bond. The probability levels for triplets are distributed mainly on 6th, 8th and 9th level whereas the probability levels for chain are distributed mainly on the 11th and 12th level.

If the probabilities were from a random source, average probability level and its standard deviation would be 6.45 and 3.45, respectively [18]. In Table 4.1, the average probability levels and standard deviations for the triplets and the chain are presented. The average probability level is 7.63 for the triplets and 9.20 for the chain whereas the standard deviations are 2.75 and 2.93, correspondingly. The raw data for the chain and triplet probability levels for each bond are given in Appendix A.3.

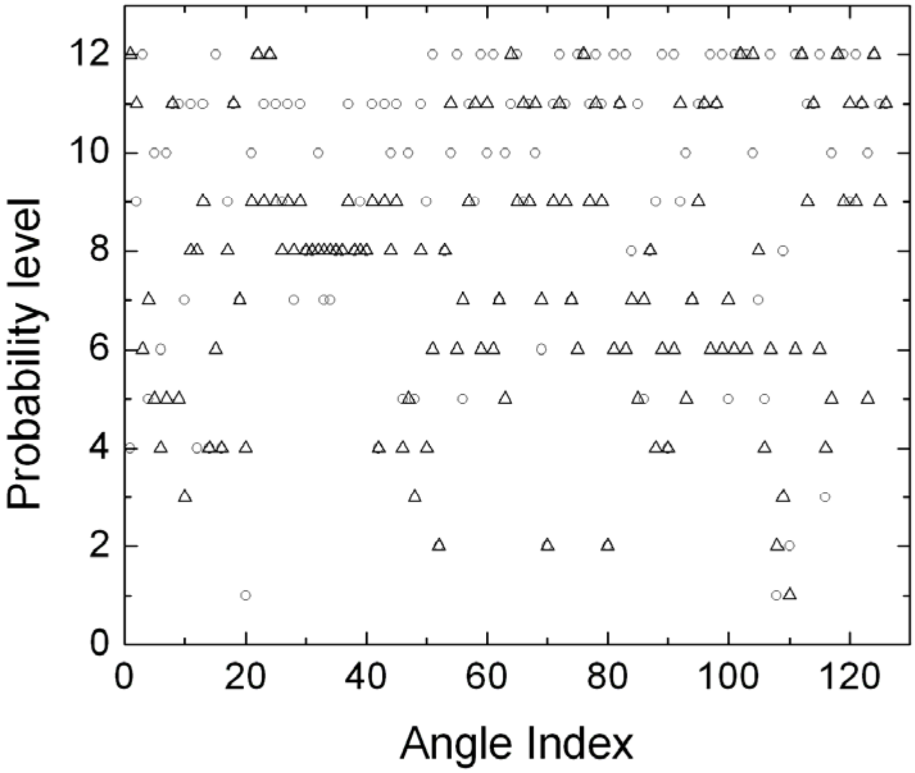

Figure 4.1. The probability levels for each bond. (For chain and triplets)

Table 4.1 The statistics of probability calculations

	Random	Triplets	Chain
Average Probability Level	6.45	7.63	9.20
Standard Deviation	3.45	2.75	2.93

The increase in the average probability levels is expected as in RIS model because of the interdependence of the bonds each other. In Figure 4.2, we try to visualize the interdependence of the bonds. Let the allowable basin for the ψ_{i-1}-φ_i map be A and the allowable basin for the φ_i -ψ_i map be B. Then the allowable basins cannot be independent. Furthermore, the allowable basin for the ψ_{i-1}-φ_i–ψ_i map can the intersection of map A and B.

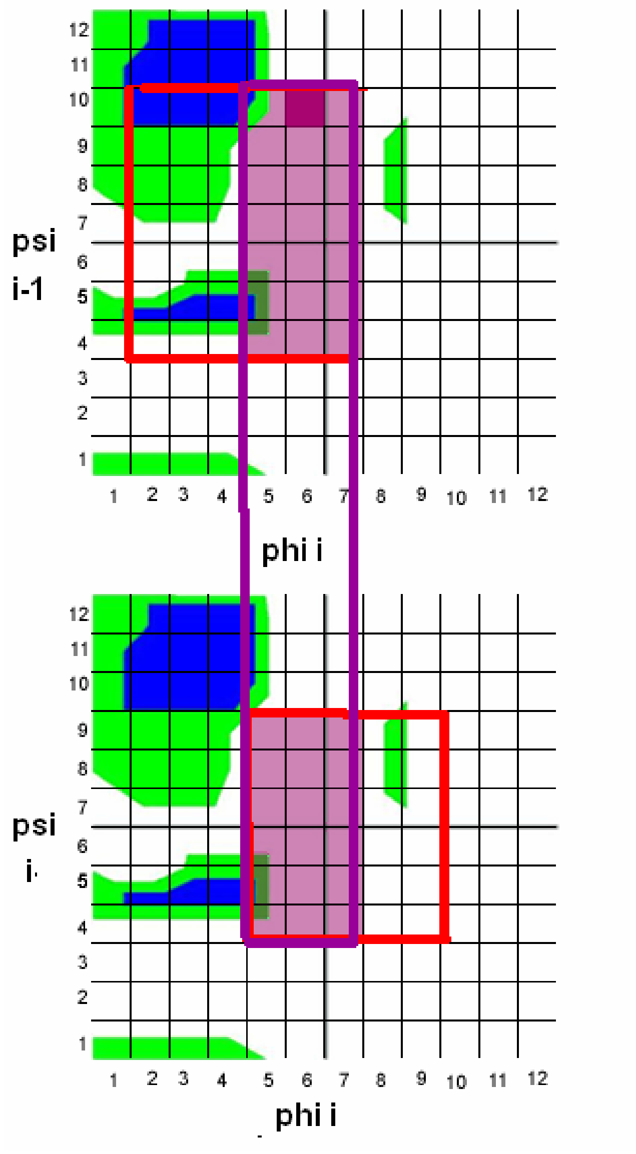

Figure4.2. Interdependence of individual bonds according to RIS model

4.2. Modifications of Stochastic Weights to Include Long-Range Effects

The average probability level, for the chain is calculated as *9.20*. These probability levels are based on the stochastic weights that reflect near neighbor interactions only, and long range effects are not included. However, the native chain conformation is obtained in the presence of long range effects also, and the probability levels should reflect long-range effects also. However, the latter effects cannot be assessed by analytical approaches, and recourse to methods such as Monte Carlo or Genetic algorithms is necessary. In this section, we try to modify the short range stochastic weights such that the calculated probability levels approach 12.0 which equate the most probable state of the chain to that of the native conformation.

4.2.1. Monte Carlo

The stochastic weights from the simulations were taken as initial stochastic weights based on short-range interactions. The method of calculation is as follows; we randomly choose one residue, and one state for its two torsion angles ϕ and ψ. Then we increase the stochastic weight of this state by 0.1*,* we recalculate bond probabilities with the new stochastic weight and check whether the average probability level of the chain increases or not. The method increases the native states' stochastic weights of randomly chosen bond with 20% probability. If the average probability level increases, we accept the change in the stochastic weight, and continue with the procedure. If the average probability level does not increase, we decline the change in the stochastic weight and continue with the procedure. The scheme for the method is shown in Figure 4.3

The highest average probability level among the 5 runs was 11.3. In order to see the difference between the different Monte Carlo simulations, we performed the statistical method, Analysis of Variance (ANOVA), which tests the difference between the means of two or more groups. The details of ANOVA are given in Appendix A.4. The results of the ANOVA test show that the difference among 5 is negligible.

However, Monte Carlo simulations were not sufficient to reach necessary highness in average probability level. Therefore we used another method to improve the average probability level. Therefore, we used another method to reflect the long-range effects on the probability levels.

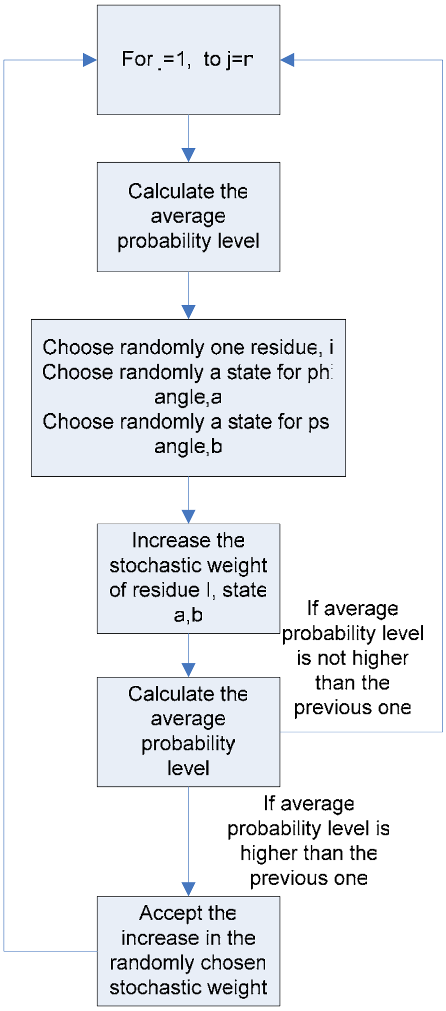
For .=1, to j=r
Calculate the average probability level
Choose randomly one residue, i
Choose randomly a state for ph: angle,a
Choose randomly a state for ps angle,b
Increase the stochastic weight of residue I, state a,b
If average probability level is not higher than the previous one
Calculate the average probability level
If average probability level is higher than the previous one
Accept the increase in the randomly chosen stochastic weight

Figure 4.3. Summary of the Monte Carlo method

4.2.2. Genetic Algorithm

Since Monte Carlo method could not reach necessary highness in probability level, we thought the problem as an optimization problem. The stochastic weights should be arranged in a way that the average probability level would be 12.

Genetic algorithms are mainly used for optimization of highly complex problems. In the present study we used GAlib, which is genetic algorithm library that contains a set of C++ genetic algorithm objects [48].

The Utilization of the Genetic Algorithm

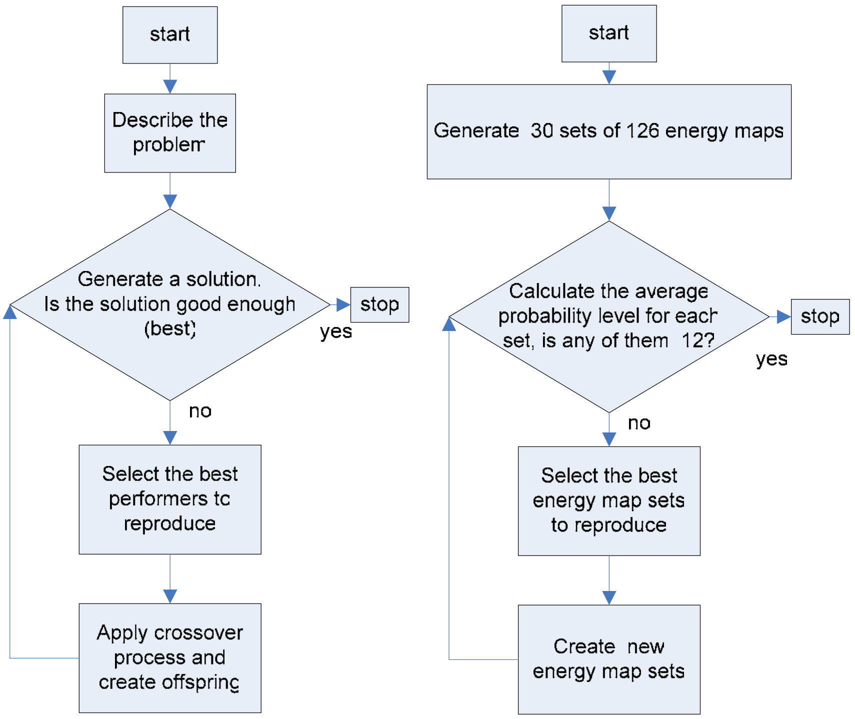
start
Describe the problem
Generate a solution. Is the solution good enough (best)
yes
stop
no
Select the best performers to reproduce
Apply crossover process and create offspring
start
Generate 30 sets of 126 energy maps
Calculate the average probability level for each set, is any of them 12?
yes
stop
no
Select the best energy map sets to reproduce
Create new energy map sets

Figure 4.4 Utilization of genetic algorithm

The outline of genetic algorithm and the application to our problem is given in Appendix A.5. The comparison between a general genetic algorithm and its application to our problem is given in Figure4.4.

Similarly to Monte Carlo simulations, we performed 5 different runs of genetic algorithm and the results of the ANOVA test shows that the differences among these 5 runs are insignificant. The highest probability level reached with genetic algorithm was 11.94.

4.3. Comparison of the Modification of the Stochastic Weights Methods

In order to see the modification of the stochastic weights better, we introduce Figure 4.5, Figure 4.6 and Figure 4.7 below. These figures are drawn for the same residue, 65th residue (Phenylalanine) of CI2. The Figure 4.5 is the energy map obtained direct from the simulations. The Figure 4.6 and Figure 4.7 are the modified energy maps obtained from Monte Carlo and genetic algorithm methods, respectively.

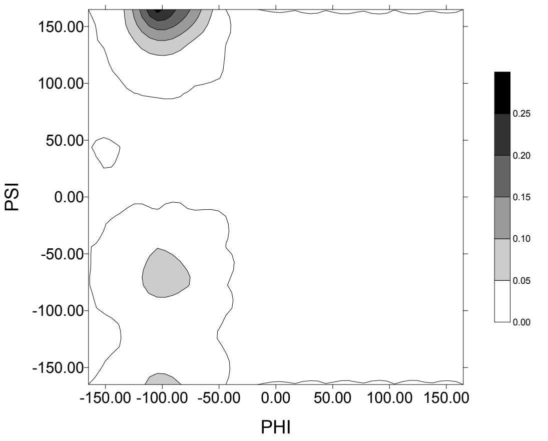

Figure 4.5 Energy map for the initial simulation result

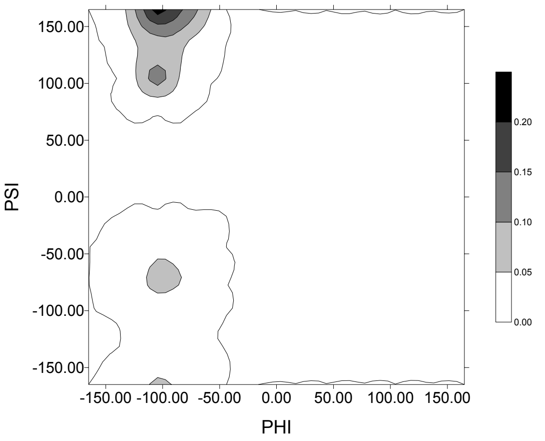

Figure 4.6 Energy map for the Monte Carlo results

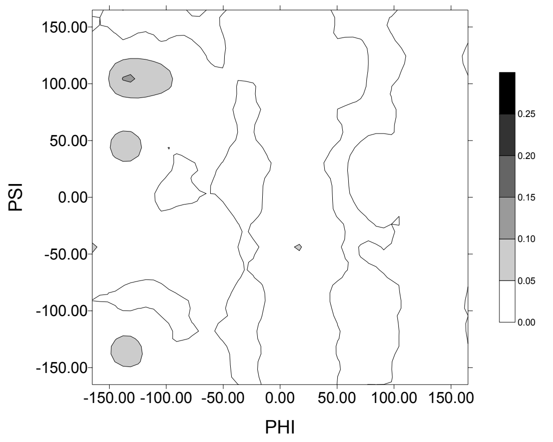

Figure 4.7 Energy map for Genetic Algorithm results

The figures are for the 65th residue (Phenylalanine) of CI2 and 65th residue is in the beta-helix structure in native CI2 conformation. The darker basins represent highness in probability.

Monte Carlo simulation provides short decrease in the probability near the alpha-helix conformation whereas a short increase in the beta-helix conformation compared to the energy map obtained from MD. However, genetic algorithm completely modifies the energy map compared to the MD simulation results. Although the conformation of the highest probability is not altered significantly, the probabilities are scattered all around the energy map obtained from the genetic algorithm compared to the energy maps obtained from MD and the Monte Carlo method.

Moreover, we compare the highest probabilities of the energy maps obtained from different methods. The highest probability value is 0.26 in Figure 4.5, 0.22 in Figure 4.6 and while 0.11 in Figure 4.7.

Any physical constraints were not given in the genetic algorithm. Although, the highest probability is in the correct configuration on the energy maps obtained from genetic algorithm, there are some basins which are physically meaningless on the energy maps.

In order to see the effects of modification of stochastic weights in details, we constructed Table 4.2, Table 4.3 and Table 4.4. The tables consist of the three lowest energies and corresponding average torsion angles. We split the results into three specific groups, which are alpha-helix, beta-sheet and turn. So, residues 39-43 are in alpha-helix conformation, residues 44-46 are in turn and residues 47-51 are in beta sheet conformation in their native states.

In Table 4.2, the three lowest energies and corresponding average torsional angles are tabulated for MD results. The lowest energies for residues in alpha helix structure are not in the native configurations; on the other hand, some of the residues' (residue 40 and residue 42) second and third lowest energies are in the native configuration. Similarly, the conformations of the lowest energies for turn structure are not predicted correctly. On the contrary, beta structures lowest energies are near the native configuration.

Table 4.2 Energies for specific secondary structures of the molecular dynamics results

	Residue	φ	Ψ	Ener(kcal)	Φ	Ψ	Ener(kcal)	φ	ψ	Ener(kcal)
Alpha-helix	residue39	165	-105	-2.31295	135	-105	-1.94453	165	-75	-1.93592
	residue40	165	-105	-2.13752	135	-105	-1.95291	-75	-105	-1.6523
	residue41	105	-105	-2.13432	135	-105	-1.94147	-45	-105	-1.43548
	residue42	-75	-105	-1.98038	165	-105	-1.81013	-105	-105	-1.70291
	residue43	-45	-75	-2.04656	-45	-105	-2.03196	165	-105	-1.98597
Turn	residue44	165	-75	-2.79078	135	-75	-2.1843	105	-75	-1.29365
	residue45	165	-105	-2.46692	-165	-105	-1.91137	165	-135	-1.52404
	residue46	-75	-105	-1.78571	165	-105	-1.7144	-75	-75	-1.65182
Beta-sheet	residue47	165	-105	-2.24117	165	-75	-1.80059	135	-105	-1.59974
	residue48	165	-105	-2.31086	165	-75	-1.94862	-75	-105	-1.7313
	residue49	165	-105	-2.12778	-75	-105	-1.86334	165	-75	-1.80302
	residue50	165	-105	-2.36787	165	-75	-2.07656	135	-105	-1.74666
	residue51	135	-105	-2.20337	165	-105	-2.0519	135	-75	-1.75661

When we study further the Monte Carlo results in Table 4.3, the configurations that give the lowest energies slightly changed. In addition, the values of the energies have slightly increased also for alpha helix, turn structure and some of residues for beta-sheet. For residues 50 and 51, the values of the lowest energies have slightly decreased.

Table 4.3 Energies for specific secondary structures of the Monte Carlo results

	Residue	φ	ψ	Ener(kcal)	Φ	Ψ	Ener(kcal)	φ	ψ	Ener(kcal)
Alpha-helix	residue39	165	-105	-2.24949	165	-75	-1.96738	135	-105	-1.8803
	residue40	135	-105	-2.1727	165	-105	-2.0149	135	-75	-1.76256
	residue41	-45	-75	-2.2753	105	-105	-1.84471	135	-105	-1.65194
	residue42	-75	-105	-2.04185	165	-105	-1.86839	-105	-105	-1.762
	residue43	-45	-75	-2.02105	-45	-105	-2.00596	165	-105	-1.95997
Turn	residue44	165	-75	-2.76378	135	-75	-2.16459	-15	-75	-1.43522
	residue45	15	-105	-2.3345	165	-105	-2.16748	-165	-105	-1.61344
	residue46	135	-105	-2.10452	-75	-105	-1.89583	165	-105	-1.85373
Beta-sheet	residue47	105	-105	-2.09815	165	-105	-1.99339	105	-135	-1.73703
	residue48	165	-105	-2.23347	165	-75	-1.94268	135	-105	-1.64921
	residue49	135	-105	-2.43539	165	-105	-1.92904	-75	-105	-1.6453
	residue50	105	-105	-2.45149	165	-105	-2.13686	165	-75	-1.6155
	residue51	165	-105	-2.42507	135	-105	-2.24299	-45	-105	-1.64019

Examining Table 4.4, we see that genetic algorithm's results predict the conformations correctly as expected. And as seen from the Figure 4.7, since the probabilities are scattered over all the configurations, lowest energies for corresponding configurations increased with respect to former results.

Table 4.4 Energies for specific secondary structures of the Genetic Algorithm results

	Residue	φ	ψ	Ener(kcal)	φ	ψ	Ener(kcal)	φ	ψ	Ener(kcal)
Alpha-helix	residue39	-45	-45	-1.89839	135	-75	-1.72992	-45	-105	-1.58499
	residue40	-15	-15	-1.6286	15	-15	-1.49675	-165	-15	-1.36397
	residue41	-45	-75	-2.4001	165	165	-2.32562	165	-135	-1.02211
	residue42	-15	-75	-1.85189	45	-75	-1.60949	-165	-75	-1.52382
	residue43	-75	-135	-1.93035	-15	-75	-1.55381	75	135	-1.38843
Turn	residue44	105	-45	-1.90135	-15	-15	-1.89875	-105	165	-1.79601
	residue45	-165	-105	-1.92087	75	165	-1.70896	15	-75	-1.47839
	residue46	135	-15	-1.83155	135	-45	-1.51646	-45	105	-1.44649
Beta-sheet	residue47	105	-165	-2.29187	-165	-165	-1.9575	-135	-105	-1.71522
	residue48	135	-75	-1.88774	-45	-45	-1.62835	-165	-165	-1.52769
	residue49	-75	-105	-2.10955	135	-165	-1.97284	135	45	-1.618
	residue50	15	165	-1.17858	-15	-105	-0.72485	-45	-105	-0.72348
	residue51	165	-135	-2.23068	165	165	-1.887	-135	165	-1.35648

The reason why MD cannot predict alpha helix conformations properly is that we constructed the stochastic weights from triplets. The formation of alpha structures depend on the hydrogen bond formation between ith and i+4 th residues [49] as we can see in Figure 4.8. We miss the $(i+3)^{rd}$ and $(i+4)^{th}$ residues in order to observe the alpha helix structure formation accurately by triplets.

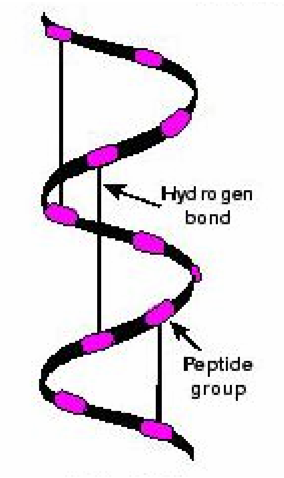

Figure 4.8. alpha helix as secondary structure adopted from [1]

To sum up and compare all the average probability level calculations results we introduce the Table 4.5. We have formerly seen that chain gives strong correlations than triplets. The same model is used by Keskin et al., but the energy maps are constructed from PDB [18]. Knowledge-based potentials give higher average probability level than MD based potentials. The characteristics of the MD are very important in this case.

The difference between average probability of triplets and chain shows that context dependence is significant in establishing the torsional bond angle preferences for the native state. Keskin et al. calculated the average probability 6.57 based on torsional angle data obtained from random configurations. Therefore one should search some factors that cause phi-psi preferences. Serrano proposed that the amount of hydrophobic surface and hydrogen-formation with the solvent could be responsible for conformational preferences. Moreover, intrinsic propensities for beta-sheet and alpha-helix point out that the chain of the amino acids determine preferences [50].

By evaluating statistical weights of torsional angles from MD, we show that even triplets have intrinsic propensities for conformational preferences. In addition, for native state of chain are context dependent, because of the significant improvement of the average probability level over chain.

Monte Carlo simulations give higher average probabilities than MD and PDB result, because we added long-range effect. Since the highest average probability level is 12, 11.34 which is the average probability level for Monte Carlo simulation is not high enough to represent the native state. The reason that Monte Carlo simulation could not reach highest average probability level may be that the system trapped into a local minimum, since we accepted only the configurations that improve the average probability level in the system. Therefore, we used one of the powerful optimization methods, genetic algorithm.

Table 4.5 Summary of comparison of the average probability levels

Method	Triplet	Chain	PDB	Monte Carlo	Genetic Algorithm
Average Probability Level	7.63	9.20	10.10	11.34	11.94

Chapter 5

STATISTICAL MECHANICS of DENATURATED PROTEINS

The basic thermodynamics and statistical thermodynamics concepts that we used in this study are given in Appendix A.6. in details. We introduced the heat capacity in terms of partition function and temperature in Appendix A.6 and we defined how to calculate partition function RIS model in chapter 3. In section 5.1, we give details of the integration of the partition function to obtain heat capacity for a denaturated protein chain.

5.1. Integration of the RIS Model Partition Function to Obtain Heat Capacity Equation

The heat capacity in terms of partition function is given in Appendix A.6:

$$C_p = \frac{2kT}{Z}\frac{\partial Z}{\partial T} - \frac{kT^2}{Z^2}\left(\frac{\partial Z}{\partial T}\right)^2 + \left(\frac{kT^2}{Z}\right)\frac{\partial^2 Z}{\partial T^2} \qquad (5.1.1)$$

The partition function, in our system is defined as the serial multiplication of stochastic weight matrices where U$_1$ and U$_n$ are the row and the column vectors. So, partition function is a scalar quantity;

$$Z = U_1 U_2 U_3 ......U_n \qquad (5.1.2)$$

The difficulty about taking the derivative of a partition function may be circumvented by introducing the term “super matrix” [47]. For simplicity, we assume that we have three stochastic weight matrices having the dimension nxn, where n is the number of states defined for the model. And the serial multiplication of the three weight matrices is

$$Z = U_1 U_2 U_3 \qquad (5.1.3)$$

The first derivative for the stochastic weight matrix is $U_1^{'} = -\frac{E_1}{R}\exp\left(-\frac{E_1}{RT}\right)$. (5.1.4)

Since U matrices are functions of $\frac{1}{T}$, Z is also function of $\frac{1}{T}$.

$$Z\left(\frac{1}{T}\right)=U_1\left(\frac{1}{T}\right)U_2\left(\frac{1}{T}\right)U_3\left(\frac{1}{T}\right) \tag{5.1.5}$$

Therefore first derivative of the partition function is defined by Equation 5.1.6

$$\frac{\partial Z}{\partial\left(\frac{1}{T}\right)}=U_1^{'}U_2U_3+U_1U_2^{'}U_3+U_1U_2U_3^{'} \tag{5.1.6}$$

By the chain rule we can rewrite $\frac{\partial Z}{\partial T}$ as in Equation 5.1.7

$$\frac{\partial Z}{\partial T}=\frac{\partial Z}{\partial\left(\frac{1}{T}\right)}\frac{\partial\left(\frac{1}{T}\right)}{\partial T}=-\frac{1}{T^2}\frac{\partial Z}{\partial\left(\frac{1}{T}\right)} \tag{5.1.7}$$

Therefore, Equation 5.1.6 can be rewritten as in Equation 5.1.8

$$\frac{\partial Z}{\partial T}=-\frac{1}{T^2}\left[U_1^{'}U_2U_3+U_1U_2^{'}U_3+U_1U_2U_3^{'}\right] \tag{5.1.8}$$

Super matrix is introduced by a 2nx2n dimensional matrix

$$A_1=\begin{bmatrix}U_1 & U_1^{'}\\ 0 & U_1\end{bmatrix} \tag{5.1.9}$$

Then serial multiplication for all super matrices is

$$A_1A_2A_3=\begin{bmatrix}U_1 & U_1^{'}\\ 0 & U_1\end{bmatrix}\begin{bmatrix}U_2 & U_2^{'}\\ 0 & U_2\end{bmatrix}\begin{bmatrix}U_3 & U_3^{'}\\ 0 & U_3\end{bmatrix}=\begin{bmatrix}U_1U_2U_3 & U_1^{'}U_2U_3+U_1U_2^{'}U_3+U_1U_2U_3^{'}\\ 0 & U_1U_2U_3\end{bmatrix}$$

(5.1.9)

and $\frac{\partial Z}{\partial\left(\frac{1}{T}\right)}$ can be extracted from the final result. The last entry of the first row or the first entry of the last column is $\frac{\partial Z}{\partial\left(\frac{1}{T}\right)}$.

Similarly the second derivative can be calculated by defining another “super matrix” [47];

$$B_1 = \begin{bmatrix} A_1 & A_1^{'} \\ 0 & A_1 \end{bmatrix} = \begin{bmatrix} U_1 & U_1^{'} & U_1^{'} & U_1^{''} \\ 0 & U_1 & 0 & U_1^{'} \\ 0 & 0 & U_1 & U_1^{'} \\ 0 & 0 & 0 & U_1 \end{bmatrix} \tag{5.1.10}$$

Serial multiplication of the B matrices is given in Equation 5.1.11;

$$B_1B_2B_3 = \begin{bmatrix} U_1 & U_1^{'} & U_1^{'} & U_1^{''} \\ 0 & U_1 & 0 & U_1^{'} \\ 0 & 0 & U_1 & U_1^{'} \\ 0 & 0 & 0 & U_1 \end{bmatrix} \begin{bmatrix} U_2 & U_2^{'} & U_2^{'} & U_2^{''} \\ 0 & U_2 & 0 & U_2^{'} \\ 0 & 0 & U_2 & U_2^{'} \\ 0 & 0 & 0 & U_2 \end{bmatrix} \begin{bmatrix} U_3 & U_3^{'} & U_3^{'} & U_3^{''} \\ 0 & U_3 & 0 & U_3^{'} \\ 0 & 0 & U_3 & U_3^{'} \\ 0 & 0 & 0 & U_3 \end{bmatrix} \tag{5.1.11}$$

$$\frac{\partial^2 Z}{\partial\left(\frac{1}{T}\right)^2} = U_1^{''}U_2U_3 + U_1^{'}U_2^{'}U_3 + U_1^{'}U_2U_3^{'} + U_1U_2^{''}U_3 + U_1^{'}U_2^{'}U_3 + U_1U_2^{'}U_3^{'} + U_1U_2U_3^{''} + U_1^{'}U_2U_3^{'} + U_1U_2^{'}U_3^{'}$$

(5.1.12)

The left hand of the Equation 5.1.12 can be extracted in the same way as in Equation 5.1.9. Formerly, we should define $\frac{\partial^2 Z}{\partial T^2}$.

$$\frac{\partial^2 Z}{\partial T^2} = \frac{\partial}{\partial T}\left(\frac{\partial Z}{\partial T}\right) = \frac{\partial}{\partial T}\left(-\frac{1}{T^2}\frac{\partial Z}{\partial\left(\frac{1}{T}\right)}\right) = -\frac{1}{T^2}\frac{\partial}{\partial\left(\frac{1}{T}\right)}\left(-\frac{1}{T^2}\frac{\partial Z}{\partial\left(\frac{1}{T}\right)}\right) = -\frac{1}{T^2}\left[-\frac{2}{T}\frac{\partial Z}{\partial\left(\frac{1}{T}\right)} - \frac{1}{T^2}\frac{\partial^2 Z}{\partial\left(\frac{1}{T}\right)}\right]$$

(5.1.13)

If we rearrange the Equation 5.1.13;

$$\frac{\partial^2 Z}{\partial T^2} = \frac{2}{T^3}\frac{\partial Z}{\partial\left(\frac{1}{T}\right)} + \frac{1}{T^4}\frac{\partial^2 Z}{\partial\left(\frac{1}{T}\right)^2} \tag{5.1.14}$$

Finally, we should insert the first and the second derivative of the partition function in to the heat capacity Equation (5.1.1) by denoting $\frac{\partial Z}{\partial T} = Z'$ and $\frac{\partial^2 Z}{\partial T^2} = Z''$.

$$C_p = \frac{2kT}{Z}\left[-\frac{1}{T^2}Z'\right] - \frac{kT^2}{Z^2}\left(\frac{1}{T^4}Z'\right)^2 + \left(\frac{kT^2}{Z}\right)\left(\frac{2}{T^3}Z' + \frac{1}{T^4}Z''\right) \tag{5.1.15}$$

Here, as the reader will notice, there are two terms cancel each other.

$$C_p = \frac{2k}{ZT}Z' - \frac{k}{Z^2T^2}(Z')^2 + \frac{2k}{ZT}Z' + \frac{k}{ZT^2}Z'' \qquad (5.1.16)$$

And the final equation for C_p becomes as in Equation 5.1.17;

$$C_p = -\frac{k}{Z^2T^2}(Z')^2 + \frac{k}{ZT^2}Z'' \qquad (5.1.17)$$

If we rearrange Equation 5.1.17 in terms of $\beta = \frac{1}{kT}$, where k is the Boltzmann constant, C_p becomes as in Equation 5.1.18

$$C_p = -\frac{1}{Z^2}\frac{1}{\beta}\frac{1}{T^3}(Z')^2 + \frac{1}{Z}\frac{1}{\beta}\frac{1}{T^3}Z'' \qquad (5.1.18)$$

5.2. Results of Heat Capacity Calculations

In addition to being one of the most important properties of a material, heat capacity is also the distinguishing characteristic of a phase transition (sudden change in one or more physical properties) [51].

In Chapter 4, we introduced three different sets of stochastic weights, each of them obtained from different methods which are MD, Monte Carlo and genetic algorithm. Heat capacity of CI2 is calculated for the three different methods as a function of temperature. The results are shown in Figure 5.1. We observe a sudden decrease in the heat capacity of CI2 for all of the three methods with increasing temperature. A sudden alteration of the heat capacity may be a consequence of the transition state which is defined as the energy barrier for proteins and transition state of CI2 was studied both experimentally and theoretically [19, 25, 52-54]. Moreover, the peak point of the heat capacity obtained from the genetic algorithm is at 298 K and much higher than that of the MD and Monte Carlo. The reason for that may be the modifications in the stochastic weights. These modifications refer to the long-range interactions.

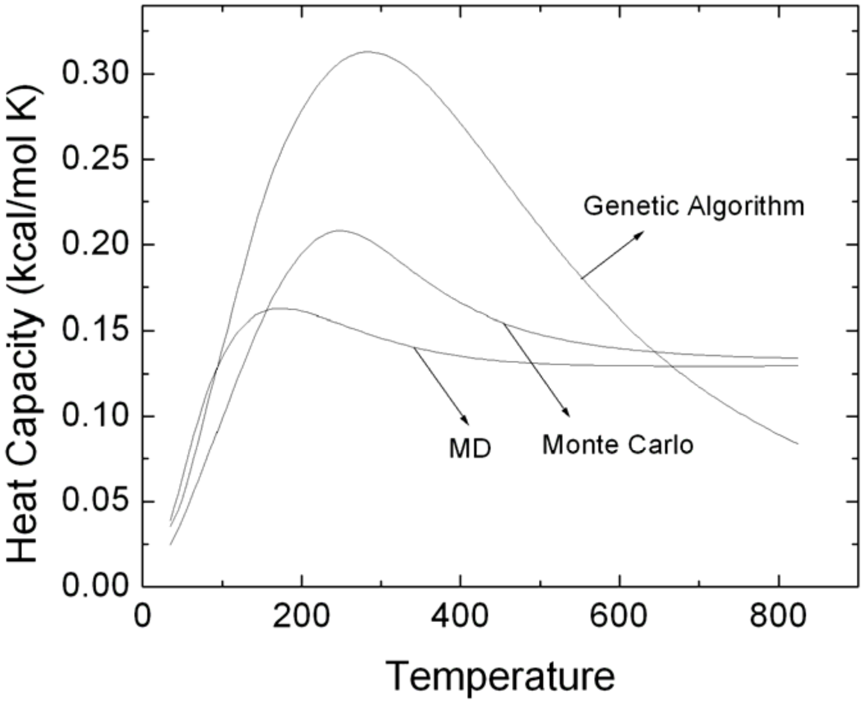

Figure 5.1. Heat capacities for different configurations

The heat capacity change of CI2 (from configuration) over temperature in aqueous solution was given by Makhatadze et al. and Table 5.1 is adopted from their work. Experimental procedures for thermodynamic properties can only specify two stable macroscopic states, the native and the denatured, and the determination of these properties depends on the temperature, pH, salt concentration [55]. The heat capacities increase for both natured and denatured states and denatured state has higher heat capacity than natured state. Entropy is higher in denatured state because configuration and non-polar contacts with water.

Makhatadze et al. stated that the entropy of protein unfolding in aqueous media includes two components: one is associated with the increase of configurational freedom in the polypeptide chain and the other with the hydration of groups that become exposed on unfolding. The configurational entropy of protein unfolding relates to the entropy changes in the absence of solvent, i.e. vacuum [56]. The configurational entropy is used several theoretical studies by assuming to have value 15-20 J / mol K per residue at 25 ^{o}C [57-59]. We defined the heat capacity in terms of temperature and partition function which contain only configuration information of CI2.

Table 5.1 Heat capacity of CI2 in aqueous Solution at

Temperature (K)	**278**	**298**	**323**	**358**	**373**	**398**
Heat Capacity native(kJ/ mol K)	12.4	13.3	14.5	15.6	16.7	17.8
Heat Capacity denaturated (kJ/ mol K)	15.8	16.9	17.8	18.0	18.2	18.0
Heat Capacity change (kJ/ mol K)	3.4	3.6	3.3	2.4	1.5	0.2

We also computed heat capacities for polyglycine obtained from MD and genetic algorithm as function of temperature in Figure 5.2. The heat capacity of polyglycine obtained from genetic algorithm is higher than the heat capacity of polyglycine obtained from MD.

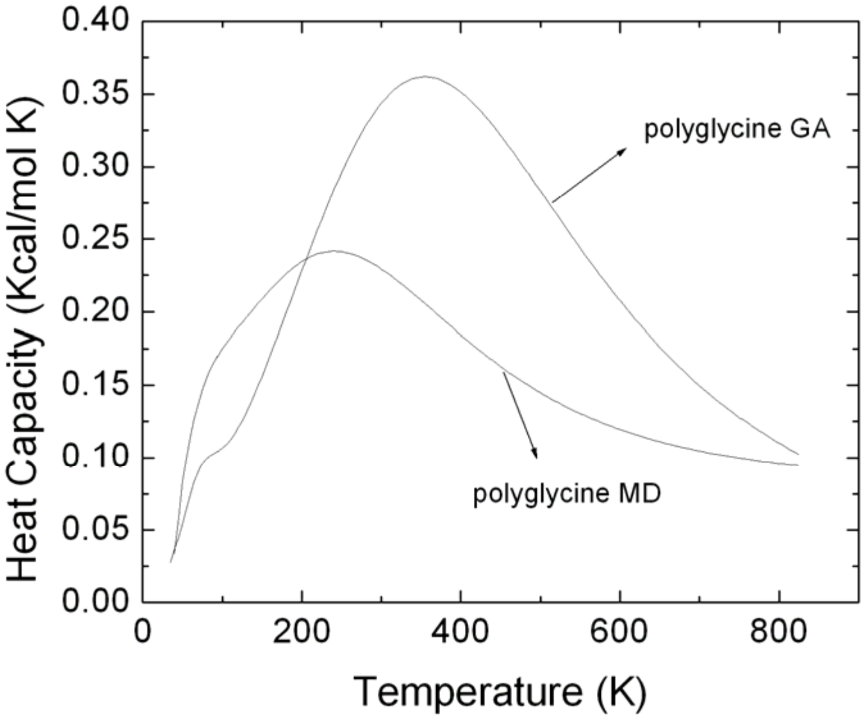

Figure 5.2. Heat capacities for polyglycine

Day et al. indicated in their study that the structure of the transition state CI2 does not change significantly under varying denaturation conditions, such as temperature [25]. At this point, one should search for a quantity that characterizes a configuration of polypeptide chain and this quantity should give information about the transition state of a protein. Radius of gyration is defined as the root-mean-square distance of the collection of atoms, from their common center of gravity. In other words, packed forms of chains have smaller radius of gyration values compared to unpacked and long chains. Transition state in terms of radius of gyration may be defined as a presence of an inflection point while the radius of gyration increases.

Calculation of $<s^2>$ defined as the average value of s^2 for over every possible configuration is given in Appendix A.7 in details.

Figure 5.3. shows the $<s^2>$ values obtained from the three methods which are MD, Monte Carlo and genetic algorithm similar to that of heat capacity calculations. Additionally, we calculated $<s^2>$ values for polyglycine that are adopted from MD and genetic algorithm.

$<s^2>$ values for polyglycine obtained from genetic algorithm and MD do not vary remarkably with respect to temperature. Moreover, $<s^2>$ values of CI2 obtained the MD and the Monte Carlo configurations increase with the temperature, however we do not observe a sudden increase. $<s^2>$ value for CI2 which configuration obtained from genetic algorithm demonstrates an inflection point as temperature increases. The inflection point of that $<s^2>$ value is at the 285 K where the heat capacity shows the peak point. In order to see the inflection point better, Figure 5.4. is drawn.

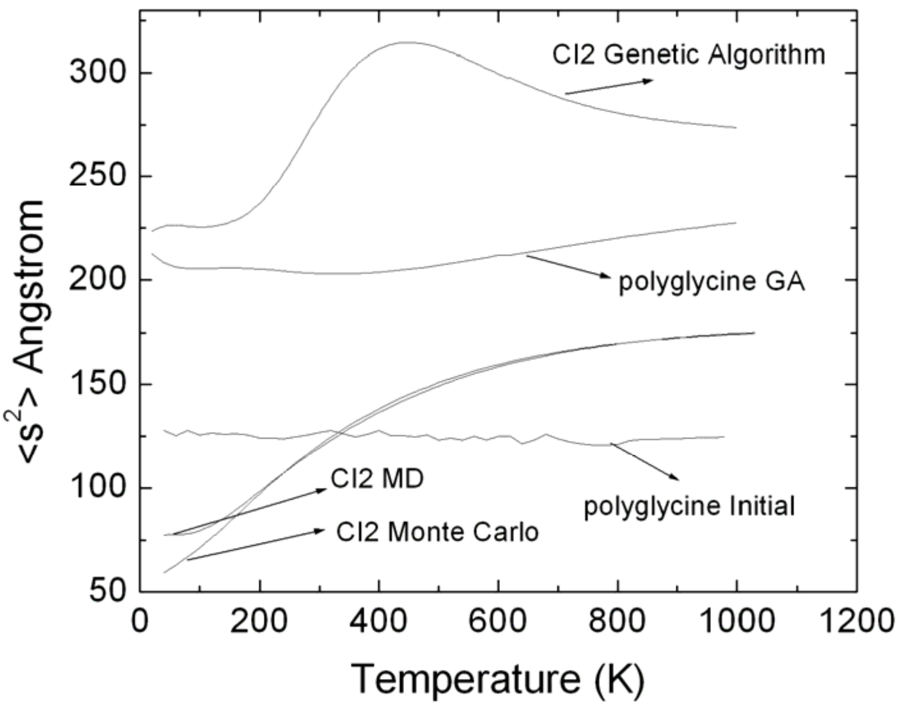

Figure5.3. $<s^2>$ values for different conformations

The radius of the gyration of native CI2 is 125 Å, but we defined the radius of gyration in terms of an average value. Therefore, we can use the term characteristic ratio which shows the departure from a freely jointed chain and every polymer has its own characteristic ratio. Miller et al. presented in their work theoretical results of average dimensions of random coil polypeptide copolymers, one of them is polyglycine [60].

The characteristic ratio is defined as the ratio of mean squared end to end bond length and number of bonds and the details are given in Appendix A.7. The characteristic ratio is calculated as two for 100% glycine content of a random coil chain [60].

When we use the polyglycine's radius of gyration value obtained from genetic algorithm, we assume 225Å because the radius of gyration slightly changes with the temperature and freely jointed chain. The characteristic ratio of this polyglycine was calculated as $1.75 \approx 2.00$.

The derivative of <s^2> values are drawn in Figure 5.4. For CI2, we observe collapse transition which is defined as the transition from an open coil form to a compact form. However, for polyglycine we do not observe a collapse transition.

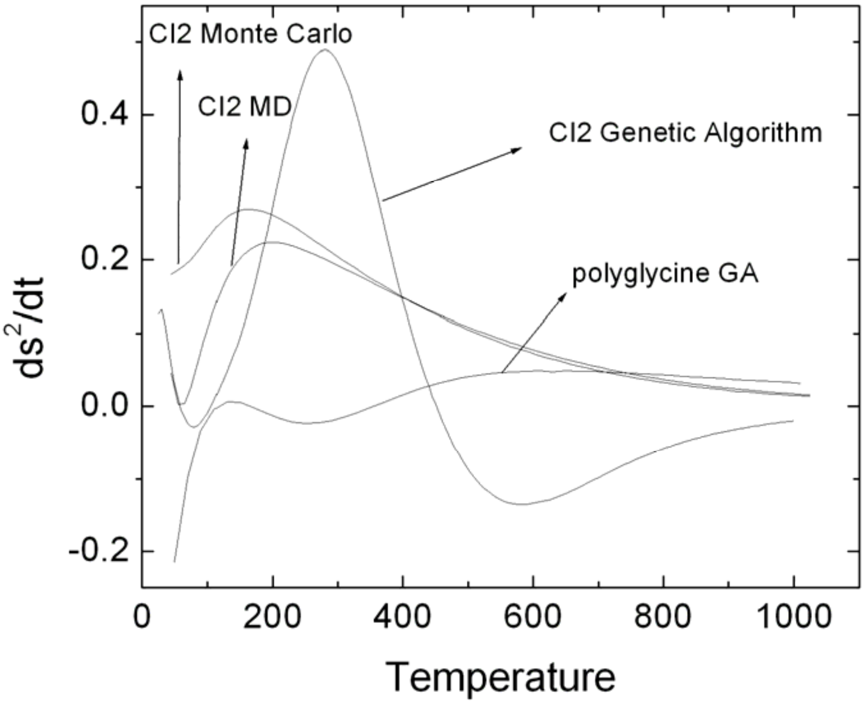

Figure 5.4. Derivative of the <s^2> for genetic algorithm

Chapter 6

LOCAL DYNAMICS OF DENATURATED PROTEINS

It is assumed in the rotational isomeric state theory that each bond can be at certain states, which are defined as angular rotations. In the present study, we define for each bond twelve different states as 30^0 intervals. And the transitions between those states are calculated from the Langevin dynamics. The simulations are described in the materials and method section. The DRIS model is analogous to RIS mode and the probability of transitions between two states is calculated. The details of the calculations are given in the Appendix A.8.

6.1. Dynamic Rotational Isomeric State Model Results

We have defined 12 discrete states that represent the torsional angle of CI2. So, we have 12^{12} transition state probabilities that are very hard to analyze. Therefore, we define four states which make transition probabilities easy to visualize. These states are commonly preferred configurations in proteins, alpha helix, beta-strand and turn. The filled yellow squares in Figure 6.1 are the native torsion angles of CI2. We defined the fourth state as all other states. The transition probabilities are defined as the sum of the transitions from other state to that state. For example, the transition probability for the alpha helix is defined as transitions to the alpha helix from all other defined states.

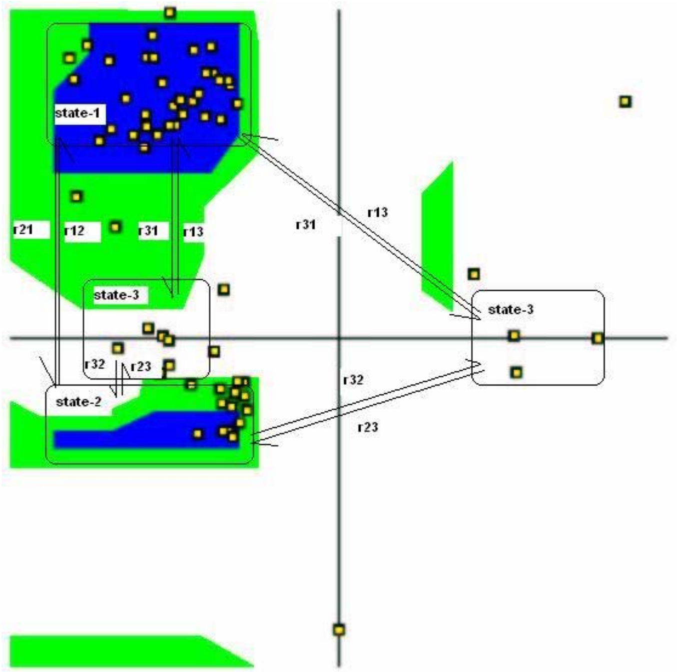

Figure 6.1 states for transition probabilities

The transition probabilities are calculated for the triplets and the chain separately. The following two figures, Figure 6.2 and Figure 6.3 represent the change in transition probabilities of the residue 69's bond (phi) as a function of time for the triplet and the chain, respectively. We can see that the governing transition probability is all-other-state whereas the transition probabilities for turn and alpha helix are stunted for the triplet. When we look at the transition probabilities for the same residues phi angle but over the chain, we can see that the transition probability for beta-strand is the leading transition probability among all other transition probabilities.

Residue 69 is in the beta-strand in native configuration of CI2. The triplet that contains residue 69 in the middle has the following residues, leucine, phenyl alanine and valine. When we look generally at the transition probabilities over the bonds that are in beta-sheets in native state of CI2, the DRIS method predicted the leading probabilities as in their native states.

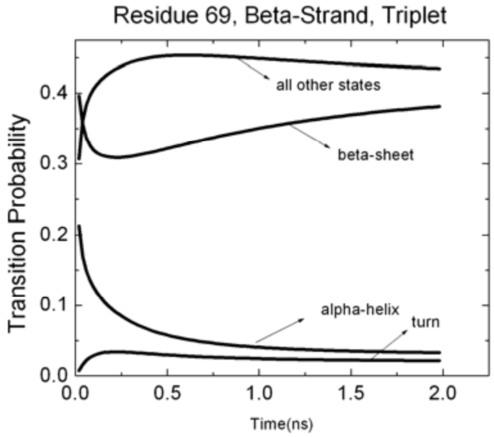

Figure 6.2. Transition probabilities of Residue 69 for triplet

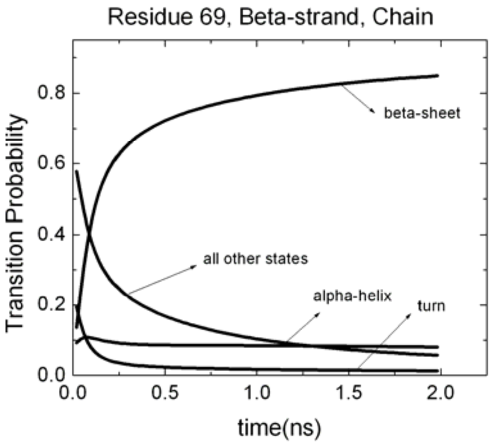

Figure 6.3. Transition probabilities of residue 69 for chain

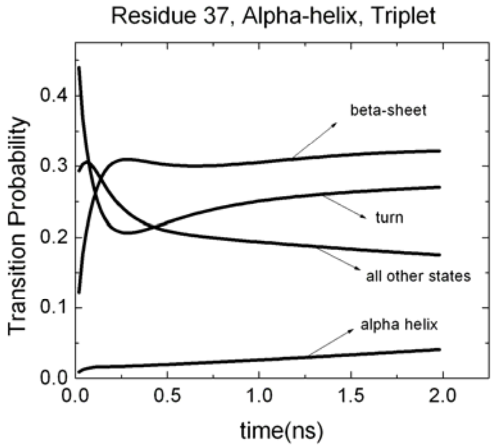

Figure 6.4 Transition probabilities of residue 35 for triplet

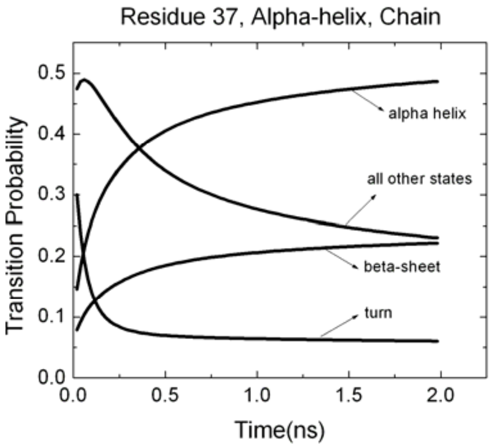

Figure 6.5 Transition probability of residue 35 for chain

Similarly, Figure 6.4 and 6.5 are the transition probabilities for residue 35 that is found in alpha helix in CI2. Figure 6.4 is for the triplet and Figure 6.5 is for the chain. In Figure 6.3, leading transition probability is for the beta-strand conformation, and the smallest transition probability is to alpha helix. The triplet itself prefers beta-strand conformation instead of alpha-helix conformation. In Figure 6.5, alpha helix becomes dominant among the transition probabilities for the chain. At t = 0 all other states has the highest transition probabilities. As time passes, the transition probability for alpha-helix dominates among all other transition probabilities. If we look generally the transition probabilities of the bonds that are found in alpha helices, we see that the leading transition probabilities are commonly predicted correctly over the chain even though the transition probabilities for triplets do not give the leading states correctly.

Similar analysis about transition probabilities for beta-strand and alpha-helix can be presented for turn. The difficulty about the turn's transition states is their definition. Turn state and alpha-helix state are very close to each other, which make it hard to distinguish the transition probabilities of the turn sate from the alpha-helix state. Among several residues which are in turn in their native state, only residue 45 has the dominant transition probability for turn. In Figure 6.6, the transition probabilities for triplets are shown for residue 45 and the dominant transition probability is for alpha helix. In fact, the transition probability for turn is very close to the transition probability for alpha helix as time increases. In Figure 6.7, the leading transition probability passed to turn from alpha-helix.

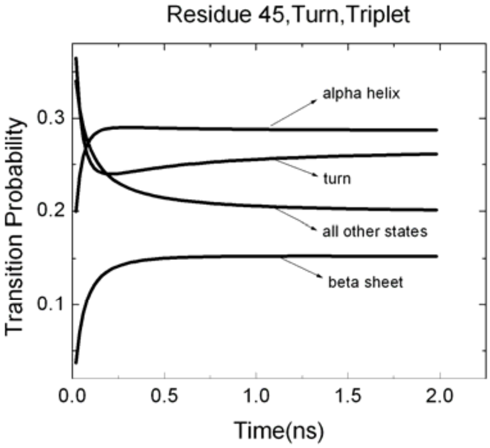

Figure 6.6 Transition probabilities of residue 45 for triplet

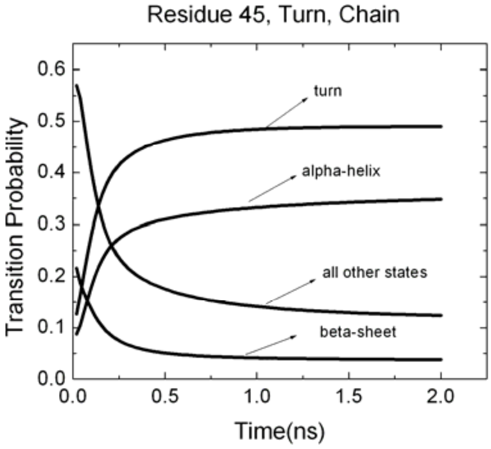

Figure 6.7 Transition probabilities of residue 45 for chain

The following analysis is the orientational autocorrelation function (OACF) that gives us a clue about the timing of the change in the transition probabilities. The details of the calculations are given in Appendix A.8. The Figures 6.8 and 6.9 are OACF as function of time for the triplet and chain. Observe that the two figures are almost identical. According to autocorrelation functions, alpha-helix formation is the fastest among other conformations. Moreover, turn formation is slower than alpha helix formation but faster than beta-strand formation.

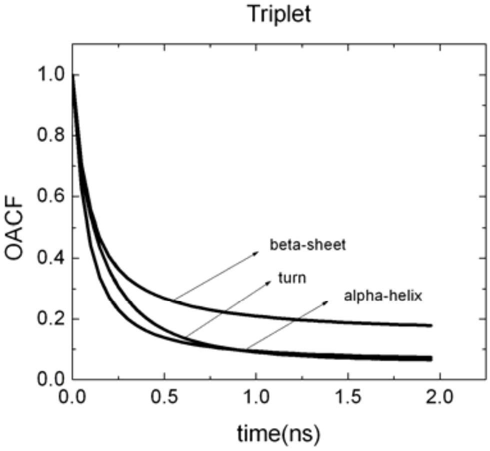

Figure 6.8 The OACF function of temperature for triplet.

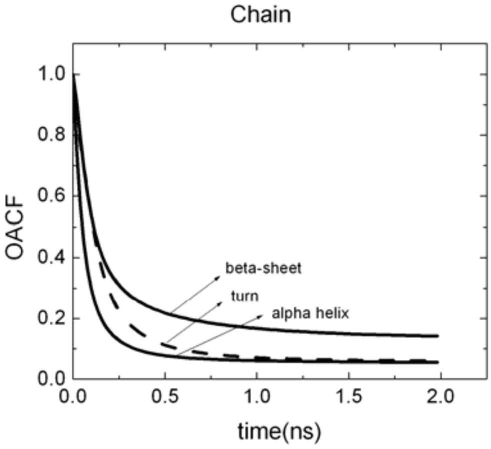

Figure 6.9 The OACF function of temperature for chain.

Chapter 7

CONCLUSION

Protein folding problem is one of the most important issues in biological sciences. There are numerous experimental and theoretical studies about investigating the effect of short and long range interactions in proteins. Understanding of preferences of torsion angle in a protein is an important approach in protein folding problem.

Our starting point involves predicting preferences of torsion angle in CI2 by using MD. CI2 has 64 residues experimentally determined. We grouped each consecutive three residues (triplets) starting from the first and performed 5 ns simulations for each triplet, starting from the 8 different initial configurations.

Although the RIS model was used for polymeric chains, Keskin et al. [18] introduced the proper way of representing stochastic weights of a polypeptide chain via knowledge-based potentials. In this study, we derived the stochastic weights from MD and evaluated them via RIS over the chain. Triplets showed different secondary structural preferences. Due to interdependency of the bonds, these preferences favor the choice of the native state torsional angles for the chain of CI2 and they are context dependent, determined by the amino acid sequence of the protein.

The model contains only short range interactions and two approaches were applied in order to include long range interaction to stochastic weights: Monte Carlo and genetic algorithms. The highest average probability level was obtained from the genetic algorithm method. Comparison of stochastic weights obtained from different methods showed that this method modifies the stochastic weights remarkably from MD and Monte Carlo results.

Additionally, the expression for the heat capacity, C_p, was derived in terms of the RIS model partition function and temperature by applying statistical mechanics concepts. Presence of transition phase was observed as a peak point in heat capacity versus temperature. For CI2, the peak coincides with the inflection point of the radius of gyration versus temperature curve.

Furthermore, local chain dynamics of CI2 was investigated via DRIS model that is analogous to the RIS model. The transition rates were derived from MD, and transition probabilities from one state to the other states were calculated for the triplets and chain separately. The differences, which were observed in the transition probabilities for residues in a alpha-helix, a beta strand and a turn. According to OACF, alpha-helix formation was the fastest among other conformations whereas turn formation was slower than alpha helix formation but faster than beta-strand formation.

The RIS model is used actually for polymer chains calculations [47, 61]. Although the RIS model is used for polymer chains, the method can be easily applied to polypeptide chains. Dill et al. concluded that proteins are polymers, therefore theories and models of polymers can be used as starting point for treating proteins [62]. Consequently, computational calculations applied to polymer chains by RIS model may be applied to proteins.

A.1 MODIFICATION IN THE TOPOLOGY FILE

```
PRES CTH3        0.00
ATOM C    CT3    -0.27
ATOM HC1  HA      0.09
ATOM HC2  HA      0.09
ATOM HC3  HA      0.09
DELETE ATOM O
BOND  C HC1 C HC2 C HC3
IC HC1  C    CA   N     1.1110  0.0000  120.0000  0.0000  0.0000
IC HC2  C    CA   N     1.1110  0.0000    0.0000  0.0000  0.0000
IC HC3  C    CA   N     1.1110  0.0000 -120.0000  0.0000  0.0000
```

A.2 CALCULATIONS OF PROBABILITY

The multiplication part can be obtained by matrix multiplication. Flory stated following equations for this purpose,

$$Z = J^* \left[\prod_{i=1}^{n} U_i \right] J \tag{A.2.1}$$

where n is the number of bonds along the chain , $J^* = \begin{bmatrix} 1 & 1 & . & . & 1 \end{bmatrix}$ and $J = \begin{bmatrix} 1 \\ 1 \\ . \\ . \\ 1 \end{bmatrix}$.

At this point, we should see the details of the multiplication. For demonstration here, we again assume two states for each bond, and create stochastic weight matrices as follows;

$$U_1 = \begin{bmatrix} u_{\alpha\alpha} & u_{\alpha\beta} \\ u_{\beta\alpha} & u_{\beta\beta} \end{bmatrix}_{12} \quad U_2 = \begin{bmatrix} u_{\alpha\alpha} & u_{\alpha\beta} \\ u_{\beta\alpha} & u_{\beta\beta} \end{bmatrix}_{12} \quad U_3 = \begin{bmatrix} u_{\alpha\alpha} & u_{\alpha\beta} \\ u_{\beta\alpha} & u_{\beta\beta} \end{bmatrix}_{12} \quad \text{....etc.}$$

$$Z = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} u_{\alpha\alpha,12}u_{\alpha\alpha,23}u_{\alpha\alpha,34}\cdots + u_{\alpha\beta,12}u_{\beta\alpha,23}u_{\alpha\alpha,34}\cdots\cdot & u_{\alpha\alpha,12}u_{\alpha\beta,23}u_{\beta\alpha,34}\cdots\cdot + u_{\alpha\beta,12}u_{\beta\beta,23}u_{\beta\alpha,34}\cdots\cdot \\ u_{\beta\alpha,12}u_{\alpha\alpha,23}u_{\alpha\beta,34}\cdots + u_{\beta\beta,12}u_{\beta\alpha,23}u_{\alpha\beta,34}\cdots & u_{\beta\alpha,12}u_{\alpha\beta,23}u_{\beta\beta,34}\cdots + u_{\beta\beta,12}u_{\beta\beta,23}u_{\beta\beta,34}\cdot\cdot \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

(A.2.2)

In this present study, we have statistical weight matrices sequences for bonds from the triplets' simulations i.e;U$_2$ U$_3$ U$_4$ for the first triplet and U$_4$ U$_5$ U$_6$ for the second triplet and etc. We have 126 bonds total and the bonds are put in order as φ_1 ψ_2 φ_3 ψ_4 etc. Each triplet is represented by 4 bonds because φ angle is not defined for the first residue of the triplet and ψ angle is not defined for the third residue of the triplet. We supply the chain connectivity by serial multiplication of last two stochastic weight matrices for each triplet. i.e.: U$_3$ U$_4$ U$_5$ U$_6$ U$_7$ U$_8$ U$_9$ U$_{10}$ etc...

The pairwise dependent probabilities are $P_{XYZ}(\phi_i, \psi_i)$ and $P_{XYZ}(\psi_i, \phi_{i+1})$ calculated as

$$P_{XYZ}(\phi_i, \psi_i) = \frac{N_{XYZ}(\phi_i, \psi_i)}{\sum N_{XYZ}} \tag{A.2.3}$$

$$P_{XYZ}(\psi_i, \phi_{i+1}) = \frac{N_{XYZ}(\psi_i, \phi_{i+1})}{\sum N_{XYZ}} \tag{A.2.4}$$

where $N_{XYZ}(\psi_i, \phi_{i+1})$ is the total number of observed conformations in that state and $\sum N_{XYZ}$ is the total number of conformations. $P_{XYZ}(\phi_i, \psi_i)$ is the probability that represents the intraresidue correlations for middle residue Y and $P_{XYZ}(\psi_i, \phi_{i+1})$ is the probability that interresidue correlations between ψ_i of residue Y and ϕ_{i+1} of residue Z. The conformational energy is defined as follows;

$$E_{XYZ}(\phi_i, \psi_i) = -RT \ln\left[\frac{P_{XYZ}(\phi_i, \psi_i)}{P^0_{XYZ}(\phi_i) P^0_{XYZ}(\psi_i)}\right] \tag{A.2.5}$$

where $P^0_{XYZ}(\phi_i)$ and $P^0_{XYZ}(\psi_i)$ are the uniform distribution probabilities, which are valid when all angles have equal probabilities. In discrete state formalism, they are directly proportional to the size of the angular intervals of the states, 12 in this case.

The statistical weights are calculated from the conformational energies. The statistical weight for bonds *i-1* and *i* in the state $\zeta\eta$ is given as:

$$u_{\zeta\eta;i} = \exp(-\frac{E_{\zeta\eta;i-1,i}}{RT}) \tag{A.2.6}$$

The probability $p_{\zeta\eta;i-1,i}$ that represents bond i-1 is in state ζ while the bond i in state η is calculated by equating all the entries to zero except the stochastic weight in state $\zeta\eta$;

$$p_{\zeta\eta;i-1,i} = Z^{-1} J^* \left[\prod_{m=1}^{i-1} U_m\right] U'_{\zeta\eta;i} \left[\prod_{m=i+1}^{n} U_m\right] J \tag{A.2.7}$$

By equating all the entries to zero, we multiply $\zeta\eta$ th state over the all possible configurations and calculate the statistical weight of that configuration $\Omega_{(\zeta\eta,i)}$. When we divide the statistical weight of that conformation to configuration partition function we obtain the probability of ζη th state for that bond.

A.3 RAW DATA FOR AVERAGE PROBABILITY LEVEL CALCULATIONS

AI	TPL	CPL	MCPL	GAPL	AI	TPL	CPL	MCPL	GAPL	AI	TPL	CPL	MCPL	GAPL
1	12	4	12	12	43	9	11	12	12	85	5	11	11	11
2	11	9	12	12	44	8	10	12	11	86	7	5	12	12
3	6	12	12	12	45	9	11	11	12	87	8	8	8	12
4	7	5	11	12	46	4	5	5	12	88	4	9	12	12
5	5	10	10	12	47	5	10	10	12	89	6	12	12	12
6	4	6	9	12	48	3	5	8	12	90	4	4	12	12
7	5	10	10	12	49	8	11	11	12	91	6	12	12	12
8	11	11	11	12	50	4	9	10	12	92	11	9	9	12
9	5	11	12	12	51	6	12	12	12	93	5	10	10	12
10	3	7	7	12	52	2	2	12	12	94	7	7	7	12
11	8	11	11	12	53	8	8	8	12	95	9	11	11	12
12	8	4	11	12	54	11	10	12	12	96	11	11	12	12
13	9	11	12	11	55	6	12	12	11	97	6	12	12	12
14	4	4	12	12	56	7	5	12	12	98	11	11	12	12
15	6	12	12	12	57	9	11	11	12	99	6	12	12	12
16	4	4	12	12	58	11	9	11	12	100	7	5	12	12
17	8	9	9	12	59	6	12	12	12	101	6	12	12	12
18	11	11	12	11	60	11	10	12	12	102	12	12	12	12
19	7	7	12	12	61	6	12	12	12	103	6	12	12	12
20	4	1	12	12	62	7	7	12	12	104	12	10	12	12
21	9	10	12	12	63	5	10	10	12	105	8	7	10	12
22	12	12	12	12	64	12	11	12	12	106	4	5	12	12
23	9	11	11	12	65	9	12	12	12	107	6	12	12	12
24	12	12	12	11	66	11	9	11	12	108	2	1	10	12
25	9	11	12	12	67	9	11	11	12	109	3	8	8	12
26	8	9	12	12	68	11	10	11	12	110	1	2	11	12
27	9	11	12	12	69	7	6	12	12	111	6	12	12	12
28	8	7	12	12	70	2	2	12	12	112	12	12	12	12
29	9	11	12	12	71	9	11	12	12	113	9	11	11	12
30	8	8	12	12	72	11	12	12	12	114	11	11	11	12
31	8	8	8	12	73	9	11	12	12	115	6	12	12	12
32	8	10	12	12	74	7	7	11	12	116	4	3	12	12
33	8	7	12	12	75	6	12	12	12	117	5	10	10	12
34	8	7	12	12	76	12	12	12	12	118	12	12	12	12
35	8	8	8	12	77	9	11	11	12	119	9	12	12	12
36	8	8	12	12	78	11	12	12	12	120	11	9	10	12
37	9	11	12	12	79	9	11	12	12	121	9	12	12	12
38	8	8	11	12	80	2	2	12	12	122	11	11	11	12
39	8	9	9	12	81	6	12	12	12	123	5	10	11	12
40	8	8	10	12	82	11	11	12	12	124	12	12	12	12
41	9	11	11	11	83	6	12	12	12	125	9	11	11	12
42	4	4	4	12	84	7	8	12	12	126	11	11	11	12

A.4 ANOVA CALCULATIONS

An ANOVA (Analysis of Variance), sometimes called an F test, is closely related to the t test. The major difference is that, where the t test measures the difference between the means of two groups, an ANOVA tests the difference between the means of two or more groups. A one-way ANOVA, tests differences between groups that are only classified on one independent variable. We have 5 different sets of stochastic weights from Monte Carlo simulations and genetic algorithm runs. The ANOVA method was applied to each bond's probability distributions. For the ANOVA test, we have 5 groups and 12 objects that correspond to simulation runs and torsional states, respectively.

$$SS_{Total} = SS_{Between} + SS_{within} \quad \text{(A.4.1)}$$

$$\sum_{j=1}^{p} n_j (\bar{x}_j - \bar{x})^2 = SS_{Between} \quad \text{(A.4.2)}$$

$$\sum_{j=1}^{p} \sum_{i=1}^{n_j} (x_{ij} - x_j)^2 = SS_{within} \quad \text{(A.4.3)}$$

We calculated the sum of squares between the groups and sum of squares within the groups from the equations A.4.2 and A.4.3, respectively. $\bar{x}$ is for the mean value over all groups, n_j is the number of states and p is the number of objects.

$$Df_b = n - 1 \quad \text{(A.4.4)}$$

$$Df_w = p(n-1) \quad \text{(A.4.5)}$$

The degrees of freedom for between groups and the within groups are given in Equation A.4.4 and A.4.5, respectively. Mean square between (MSB) and mean square between (MSW) are defined as the ratio of the sum of squares to the degrees of freedom.

$$MSB = \frac{SS_{between}}{Df_b} \quad \text{(A.4.6)}$$

$$MSW = \frac{SS_{between}}{Df_w} \quad \text{(A.4.7)}$$

The F-value is defined as the ratio of the MSB and MSW. The critical value for 5 groups and 12 subjects is given as 3.15. The F values we calculated from Monte Carlo and genetic algorithm simulations are between 0 and 1, far from critical value.

.

A.5 GENETIC ALGORITHM

Basic Description of Genetic Algorithm

Genetic algorithms are inspired by Darwin's theory of evolution. Solution to a problem by genetic algorithms uses an evolutionary process. The algorithm begins with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population. This is motivated by a hope, that the new population will be better than the old one. Solutions which are then selected to form new solutions (offspring) are selected according to their fitness - the more suitable they are the more chances they have to reproduce. This is repeated until some condition (for example number of populations or improvement of the best solution) is satisfied. The outline for basic genetic algorithm is as follows:

1. **[Start]** Generate random population of *n* chromosomes (suitable solutions for the problem)
2. **[Fitness]** Evaluate the fitness *f(x)* of each chromosome *x* in the population
3. **[New population]** Create a new population by repeating the following steps until the new population is complete
    1. **[Selection]** Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)
    2. **[Crossover]** With a crossover probability cross over the parents to form new offspring (children). If no crossover was performed, offspring is the exact copy of parents.
    3. **[Mutation]** With a mutation probability mutate new offspring at each locus (position in chromosome).
    4. **[Accepting]** Place new offspring in the new population
4. **[Replace]** Use new generated population for a further run of the algorithm
5. **[Test]** If the end condition is satisfied, **stop**, and return the best solution in current population
6. **[Loop]** Go to step **2**

As we can see, the outline of the basic genetic algorithm is very general. The first question to ask is how to create chromosomes and what type of encoding to choose. The next question is how to select parents for crossover. This can be done in many ways, but the main idea is to select the better parents (best survivors) in the hope that the better parents will produce better offspring. A chromosome should in some way contain information about solution that it represents. In our problem, we search for a complete set of stochastic weight matrices that would give the average probability level as 12 for CI2. Therefore, our chromosomes are set of stochastic weights represent the stochastic weights of bonds in CI2, i.e.: U_1 U_2 U_3 U_4 U_5 U_6 etc. After we have decided what encoding we will use, we can proceed to crossover operation but first we have to choose appropriate chromosomes to crossover. The average probability level is calculated for each chromosome (set of stochastic weights) and the chromosomes, which have higher average probability, have bigger chance to be selected. This is the key part of the evolution. Crossover can be illustrated as follows: (| is the crossover point, and assume that we have eight stochastic weight matrices for 8 bonds for demonstration) :

Chromosome 1	U_1 U_2 U_2 U_4 | U_5 U_6 U_7 U_8
Chromosome 2	U_1 U_2 U_2 U_4 | U_5 U_6 U_7 U_8
Offspring 1	U_1 U_2 U_2 U_4 | U_5 U_6 U_7 U_8
Offspring 2	U_1 U_2 U_2 U_4 | U_5 U_6 U_7 U_8

Now, we have the new population reproduced from the best parents (set of stochastic weights that give the highest average probability results). After a crossover is performed, mutation takes place. Mutation is intended to prevent falling of all solutions in the population into a local optimum of the solved problem. Mutation operation randomly changes the offspring results from crossover. As mutation, we changed the stochastic weight for randomly chosen states of a randomly chosen bond. Crossover and mutation probabilities are used as default values of GAlib [48]. Then we placed the new offspring (crossover and mutated) in the new population and we use the new population for further run of the algorithm. We run the algorithm for 100,000 generations.

A.6. STATISTICAL MECHANICS

Thermodynamic Connection

The partition function can be used to find the expected (average) value of any microscopic property of the system, which can then be related to macroscopic variables. For instance, the expected value of the microscopic energy E is *interpreted* as the microscopic definition of the thermodynamic variable internal energy (U), and can be obtained by taking the derivative of the partition function with respect to the temperature[51]

Then internal energy which is the average energy is defined as $U = -\left(\frac{\partial \ln Z}{\partial \beta}\right)_{N,V}$ (A.6.1)

Due to chain rule, internal energy can be written as;

$$\langle E \rangle = \frac{\sum_i E_i e^{-\beta E_i}}{Z} = \frac{-\frac{dZ}{d\beta}}{Z} \qquad \text{(A.6.2)}$$

$$U = -\left(\frac{\partial \ln Z}{\partial T}\frac{\partial T}{\partial \beta}\right) \qquad \text{(A.6.3)}$$

First part of the equation is;

$$\frac{\partial \ln Z}{\partial T} = \frac{1}{Z}\frac{\partial Z}{\partial T} \qquad \text{(A.6.4)}$$

Second part of the equation can be computed as follows;

T in terms of β is $T = \frac{1}{k\beta}$ (A.6.5)

and the derivative of T in terms of β is $\frac{\partial T}{\partial \beta} = -kT^2$. (A.6.6)

So, internal energy becomes,

$$U = \frac{kT^2}{Z}\frac{\partial Z}{\partial T}. \tag{A.6.6}$$

Since heat capacity is defined as follows,

$$C_p = \left(\frac{\partial U}{\partial T}\right)_P \tag{A.6.7}$$

if we insert U in to the equation above, C_p becomes;

$$C_p = \frac{\partial\left(\frac{kT^2}{Z}\frac{\partial Z}{\partial T}\right)}{\partial T} \tag{A.6.8}$$

Due to the product rule, Equation isA.5.21 becomes

$$C_p = \left(\frac{\partial\left(\frac{kT^2}{Z}\right)}{\partial T}\right)\frac{\partial Z}{\partial T} + \left(\frac{kT^2}{Z}\right)\frac{\partial^2 Z}{\partial T^2} \tag{A.6.9}$$

Finally we get the final result by computing all the necessary derivations

$$C_p = \left(\frac{2kTZ - kT^2\frac{\partial Z}{\partial T}}{Z^2}\right)\frac{\partial Z}{\partial T} + \left(\frac{kT^2}{Z}\right)\frac{\partial^2 Z}{\partial T^2} \tag{A.6.10}$$

Rearranging the Equation A.5.22 we get

$$C_p = \frac{2kT}{Z}\frac{\partial Z}{\partial T} - \frac{kT^2}{Z^2}\left(\frac{\partial Z}{\partial T}\right)^2 + \left(\frac{kT^2}{Z}\right)\frac{\partial^2 Z}{\partial T^2} \tag{A.6.11}$$

A.7 RADIUS OF GYRATION CALCULATIONS

Square of the Magnitude of the Chain Vector

Vectors are denoted by an arrow i.e. $\vec{l}$ and matrices are denoted by capital letter i.e. T

The square of the magnitude of r is given by

$$r^2 = \sum_1^n l_h^2 + 2\sum_{h<j} \vec{l}_h^{\,T} T_h T_{h+1} ....T_{j-1}\vec{l}_j \qquad (A.7.1)$$

where $\vec{l}_h^{\,T}$ is the row form of the bond vector and l_h is its magnitude and T is the transformation matrix. Transformation matrix is defined as follows;

$$T = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ \sin\theta\cos\varphi & -\cos\theta\cos\varphi & \sin\varphi \\ \sin\theta\sin\varphi & -\cos\theta\sin\varphi & -\cos\varphi \end{bmatrix} \qquad (A.7.2)$$

where θ is for the angle between the bonds φ is for the torsional angles. In the following tables, the fixed values for the angles and lengths are tabulated.

Torsional Angles	**Length (A)**
Phi	1.47
Psi	1.53
Omega	1.32

Bond Angles	**Degree**
<C$^\alpha$CN	66
< CNC$^\alpha$	57
< NC$^\alpha$C	70

For bond i the generator matrix is proposed;

$$G_i = \begin{bmatrix} 1 & 2\vec{l}^{\,T}T & l^2 \\ \vec{0} & T & \vec{l} \\ 0 & \vec{0} & 1 \end{bmatrix}_i \qquad (A.7.3)$$

The square of the magnitude of the chain vector is defined as:

$$r^2 = G_{[1} G_2^{(n-2)} G_{n]} \quad n \geq 2 \tag{A.7.4}$$

where $G_{[1}$ and $G_{n]}$ represents the first row of G_1 and the final column of G_n, respectively.

If we look closely to the Equation (A.7.4);

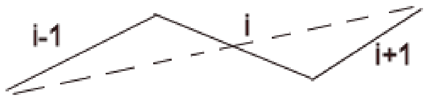

Figure A.7.1. The distance r^2 between i-1th and i+1 th bond

The dashed line is the distance (r) between the bonds i-1 and i+1. So the multiplication of the generator matrices is, as it was given in Equation (A.6.4);

$$r^2 = \begin{bmatrix} 1 & 2\vec{l}_{i-1}^{\,T} T_{i-1} & l_{i-1}^2 \end{bmatrix} \begin{bmatrix} 1 & 2\vec{l}_i^{\,T} T_i & l_i^2 \\ \vec{0} & T_i & \vec{l}_i \\ 0 & \vec{0} & 1 \end{bmatrix} \begin{bmatrix} l_{i+1}^2 \\ \vec{1} \\ 1 \end{bmatrix} \tag{A.7.5}$$

$$r^2 = l_{i-1}^2 + l_i^2 + l_{i+1}^2 + 2(\vec{l}_{i-1}^{\,T} T_{i-1} \vec{l}_i + \vec{l}_{i-1}^{\,T} T_{i-1} T_i \vec{l}_{i+1} + \vec{l}_i^{\,T} T_i \vec{l}_{i+1}) \tag{A.7.6}$$

The matrix multiplication ends up to be in the form of equation (A.7.1) which is nothing but the summation of three vectors. If we have more than one matrix in the middle, the serial multiplication of G matrices will be identical to Equation (A.7.1).

The matrix U_i represents the matrix of statistical weights $u_{\zeta\eta,i}$ applicable to rotational isomeric states $\zeta\eta = 1,2,3....n$ for bonds i-1 and i. It is given as $u_{\zeta\eta,i} = \exp\left(-\frac{E_{\zeta\eta;i-1,i}}{RT}\right)$ and $E_{\zeta\eta;i-1,i} = -RT\ln\left[\frac{P_{i-1,i}}{P_{i-1}^0 P_i^0}\right]$ is defined as conformational energy where P^0 denotes the uniform distribution probabilities.

The configuration partition function for the chain is given by the serial product of U_i matrices

$$Z = U_{i-1}U_iU_{i+1}.....U_n \tag{A.7.7}$$

When we consider a configuration-dependent molecular property, $f = f(\{\Phi\})$, we assume that this property can be stated as the sum of the contributions of each individual bond of the chain.

The Generator matrix is defined by

$$\Im_i = (U_i \otimes E_s)\|F_i\| \tag{A.7.8}$$

Where $\|F_i\|$ is the diagonal array of the generator matrices $F(1), F(2).......F(n)$ for the rotational states 1,2,……n and E$_s$ is the matrix identity of the order s of matrix F$_i$. The dimensions of the $\Im$ matrix is nxs, where n is the number of the isomeric states, and s is the dimension of F matrix.

Equation (A.7.8) is actually formulated as;

$$\Im_i = \begin{bmatrix} u_{11}\cdot F(1) & u_{12}\cdot F(2) & .... & .... \\ u_{21}\cdot F(1) & u_{22}\cdot F(2) & .... & .... \\ .... & .... & ..... & .... \\ .... & .... & .... & u_{nn}\cdot F(n) \end{bmatrix} \tag{A.7.9}$$

The reader should figure out that serial multiplication of $\Im_i$ matrices simultaneously generates the product of the statistical weights $u_{\zeta\eta}$'s and every statistical weight is multiplied with the generator matrix $F(\eta)$ for the same configuration. Therefore, the serial multiplication of $\Im_i$ ‘s includes the sum of the complete set of statistical weights and generator matrices for each configuration. Division by the sum Z yields the average of <f>.

Statistical Mechanical Averages over the Configurations

In detailed form of the matrices, please observe the difference, assuming there are three rotational isomeric states;

$$Z = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} & & \\ & & \\ u_{\omega\varphi,1} & u_{\omega\varphi,2} & u_{\omega\varphi,3} \end{bmatrix} \begin{bmatrix} u_{\varphi\psi,11} & u_{\varphi\psi,12} & u_{\varphi\psi,13} \\ u_{\varphi\psi,21} & u_{\varphi\psi,22} & u_{\varphi\psi,23} \\ u_{\varphi\psi,31} & u_{\varphi\psi,32} & u_{\varphi\psi,33} \end{bmatrix} \begin{bmatrix} & & u_{\psi\omega,1} \\ & & u_{\psi\omega,2} \\ & & u_{\psi\omega,3} \end{bmatrix} \cdots\cdots\cdots \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}. \qquad \text{(A.7.10)}$$

The third matrix that represent ω is given as identity, because it is fixed to an angle and doesn't rotate.

$$Z = u_{\varphi,1} \cdot u_{\varphi\psi,11} \cdot 1_{\omega} + u_{\varphi,1} \cdot u_{\varphi\psi,12} \cdot 1_{\omega} + u_{\varphi,1} \cdot u_{\varphi\psi,13} \cdot 1_{\omega} + u_{\varphi,2} \cdot u_{\varphi\psi,21} \cdot 1_{\omega} + u_{\varphi,2} \cdot u_{\varphi\psi,22} \cdot 1_{\omega} + u_{\varphi,2} \cdot u_{\varphi\psi,23} \cdot 1_{\omega}$$

$$+ u_{\varphi,3} \cdot u_{\varphi\psi,31} \cdot 1_{\omega} + u_{\varphi,3} \cdot u_{\varphi\psi,32} \cdot 1_{\omega} + u_{\varphi,3} \cdot u_{\varphi\psi,33} \cdot 1_{\omega} \quad \text{(A.7.11)}$$

The major difference comes up with the angle omega, because omega angle can be only at one state. When defining the generator matrix $\Im_i$ for the present case; $\Im_\omega$ should be treated differently from the other two matrices, because omega torsional angle is fixed to state 3.

$$\Im_\varphi = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ u_{\omega\varphi,1} \cdot F(1) & u_{\omega\varphi,2} \cdot F(2) & u_{\omega\varphi,3} \cdot F(3) \end{bmatrix} \qquad \text{(A.7.12)}$$

$$\Im_{\varphi\psi} = \begin{bmatrix} u_{\varphi\psi,11} \cdot F(1) & u_{\varphi\psi,12} \cdot F(2) & u_{\varphi\psi,13} \cdot F(3) \\ u_{\varphi\psi,21} \cdot F(1) & u_{\varphi\psi,22} \cdot F(2) & u_{\varphi\psi,23} \cdot F(3) \\ u_{\varphi\psi,31} \cdot F(1) & u_{\varphi\psi,32} \cdot F(2) & u_{\varphi\psi,33} \cdot F(3) \end{bmatrix} \qquad \text{(A.7.13)}$$

$$\Im_\omega = \begin{bmatrix} 0 & 0 & u_{\psi\omega,1} \cdot F(3) \\ 0 & 0 & u_{\psi\omega,2} \cdot F(3) \\ 0 & 0 & u_{\psi\omega,3} \cdot F(3) \end{bmatrix} \qquad \text{(A.7.14)}$$

Each $\Im$ matrix has the dimension of 3x5, where 3 come from the rotational isomeric states and 5 from the generator matrix's dimension. The serial multiplication of $\Im$ matrices will give the following result;

$$
\begin{aligned}
&u_{\varphi,1}\cdot F(1)\cdot u_{\varphi\psi,11}\cdot F(1)\cdot 1_{\omega}\cdot F(3)+u_{\varphi,1}\cdot F(1)\cdot u_{\varphi\psi,12}\cdot F(2)\cdot 1_{\omega}\cdot F(3)+u_{\varphi,1}\cdot F(1)\cdot u_{\varphi\psi,13}\cdot F(3)\cdot 1_{\omega}\cdot F(3)\\
&+u_{\varphi,2}\cdot F(2)\cdot u_{\varphi\psi,21}\cdot F(1)\cdot 1_{\omega}\cdot F(3)+u_{\varphi,2}\cdot F(2)\cdot u_{\varphi\psi,22}\cdot F(2)\cdot 1_{\omega}\cdot F(3)+u_{\varphi,2}\cdot F(2)\cdot u_{\varphi\psi,23}\cdot F(3)\cdot 1_{\omega}\cdot F(3)\\
&+u_{\varphi,3}\cdot F(3)\cdot u_{\varphi\psi,31}\cdot F(1)\cdot 1_{\omega}\cdot F(3)+u_{\varphi,3}\cdot F(3)\cdot u_{\varphi\psi,32}\cdot F(2)\cdot 1_{\omega}\cdot F(3)+u_{\varphi,3}\cdot F(3)\cdot u_{\varphi\psi,33}\cdot F(3)\cdot 1_{\omega}\cdot F(3)
\end{aligned}
$$

Finally <f> is defined as

$$
<f>=\frac{\Im_{[1}\Im_{2}\Im_{3}\Im_{4}.......\Im_{n]}}{Z} \qquad \text{(A.7.15)}
$$

where $\Im_{[1}$ is the first row of the $\Im_{1}$ and $\Im_{n]}$ is the last column of the $\Im_{n}$.

One should notice that first row of the generator matrix contains only 1st state for $\Im_{\varphi}$ and only one 3rd state for $\Im_{\omega}$. Therefore only for the terminal matrices they should be defined as follows while calculating the average quantity,

$$
\Im_{\varphi}=\begin{bmatrix} u_{\omega\varphi,1}\cdot F(1) & u_{\omega\varphi,2}\cdot F(2) & u_{\omega\varphi,3}\cdot F(3)\\ 0 & 0 & 0\\ 0 & 0 & 0 \end{bmatrix} \text{ if it is the left terminal}
$$

And $\Im_{\omega}=\begin{bmatrix} 0 & 0 & u_{\psi\omega,1}\cdot F(3)\\ 0 & 0 & u_{\psi\omega,2}\cdot F(3)\\ 0 & 0 & u_{\psi\omega,3}\cdot F(3) \end{bmatrix}$ if it is the right terminal

For any two bonds in the chain, this result can be easily modified to following form,

$$
<f_{hl}>=\frac{(U_{1}U_{2}....U_{h-1})\Im_{[h}\Im_{h+1}\Im_{h+2}\Im_{h+3}.......\Im_{k]}(U_{k+1}U_{k+2}....U_{n})}{Z} \qquad \text{(A.7.16)}
$$

In the specific case, we want to calculate $<r^2>$, therefore the generator matrix is G (see Appendix A.4). So the former equation becomes

$$
<r_{kl}^{2}>=\frac{(U_{1}U_{2}....U_{h-1})\Im_{[h}\Im_{h+1}\Im_{h+2}\Im_{h+3}.......\Im_{k]}(U_{k+1}U_{k+2}....U_{n})}{Z} \qquad \text{(A.7.17)}
$$

where $\Im_{i}$ is defined as

$$
\Im_{i}=\begin{bmatrix} u_{11}\cdot G(1) & u_{12}\cdot G(2) & .... & ....\\ u_{21}\cdot G(1) & u_{22}\cdot G(2) & .... & ....\\ .... & .... & ..... & ....\\ .... & .... & .... & u_{nn}\cdot G(n) \end{bmatrix} \qquad \text{(A.7.18)}
$$

Radius of gyration

Lagrange's theorem which is related the center of gravity for a system of masses to the distances between their centers taken pairwise is as follows;

s_i is the vector from the center of gravity to chain atom I, and let r_{0i} be the vector be from the zeroth atom to the ith of chain in its specified configuration.

Then $s_i=s_0+r_{0i}$ s_0 being the vector leading from the center of gravity to the zeroth atom. The square of the radius of gyration is by definition

$$s^2 = (n+1)^{-1}\sum_0^n s_i^2 = (n+1)^{-1}\sum_0^n s_i \cdot s_i \tag{A.7.19}$$

If we substitute s_i into the foregoing equation, we have

$$s^2 = (n+1)^{-1}\sum_0^n s_i^2 = (n+1)^{-1}\sum_0^n (s_0 + r_{0i})\cdot(s_0 + r_{0i}) \tag{A.7.20}$$

In more details, the equation becomes

$$s^2 = s_0^2 + 2(n+1)^{-1} s_0 \cdot \sum_1^n r_{0i} + (n+1)^{-1}\sum_1^n r_{0i} \tag{A.7.21}$$

Inasmuch as $\sum_0^n s_i = 0$ then s_0 becomes $s_0 = -(n+1)^{-1}\sum_1^n r_{0i}$ and

$$s^2 = (n+1)^{-2}\sum_{i=1}^n\sum_{j=1}^n r_{0i}\cdot r_{0j} \tag{A.7.22}$$

So s^2 turns into the following equation

$$s^2 = (n+1)^{-1}\sum_1^n r_{0i}^2 - (n+1)^{-2}\sum_{i=1}^n\sum_{j=1}^n r_{0i}\cdot r_{0j} \tag{A.7.23}$$

Due to the law of cosines

$$r_{0i}\cdot r_{oj} = \frac{r_{0i}^2 + r_{0j}^2 - r_{ij}^2}{2} \tag{A.7.24}$$

Finally we have

$$s^2 = (n+1)^{-1}\left[\sum_{i=1}^n r_{0i}^2 + \sum_{i=1}^n\sum_{j=1}^n r_{ij}^2\right] = \frac{1}{2}(n+1)^{-2}\sum_{i=1}^n\sum_{j=1}^n r_{ij}^2 = (n+1)^{-2}\sum_{0\le i<j\le n}^n r_{ij}^2 \tag{A.7.25}$$

In order to calculate radius of gyration, we calculated $< r_{ij}^2 >$ for every alpha carbon. Therefore, s^2 is the average value over the all possible conformations, $<s^2>$.

Characteristic Ratio

The relationship between $<s^2>$ and average $<r^2>$ is introduced by Flory [47] when we assume the chain is long enough: $< s^2 >= \frac{< r^2 >}{6}$ (A.7.26)

So the characteristic ratio becomes: $C_\infty = \frac{< s^2 > 6}{n_p l_p^2}$ (A.7.27)

where n_p is the number of the virtual bonds and l_p is the length of the virtual bonds.

A.8. CONFIGURATION STOCHASTICS OF CHAIN DYNAMICS

For a sequence consisting of *N* skeletal bonds the time rate of change of $P^{(N)}(t)$ is defined as follows:

$$\frac{dP^{(N)}(t)}{dt} = A^{(N)}P^{(N)}(t) \qquad \text{(A.8.1)}$$

where $A^{(N)}$ is the 12^Nx12^N matrix the elements of which describing the rate from one state to another. The solution to the equation (A.7.1) is

$$P^{(N)}(t) = \exp(A^{(N)}t)P^{(N)}(t=0) = B^{(N)}\exp\left(L^{(N)}t\right)\left[B^{(N)}\right]^{-1}P^{(N)}(t=0) \qquad \text{(A.8.2)}$$

where $B^{(N)}$ is the matrix form of eigenvectors of $A^{(N)}$ and $[B^{(N)}]^{-1}$ is the inverse of the eigenvectors matrix. Also, $L^{(N)}$ is the diagonal matrix of eigenvalues of $A^{(N)}$. So, we can define the time-dependent probability matrix as $C^{(N)}$, which is

$$C^{(N)} = B^{(N)}\exp(L^{(N)}t)(B^{(N)})^{-1} \qquad \text{(A.8.3)}$$

We define the equation for two bonds by $P^{(2)}(t) = C^{(2)}diagP^{(2)}(t=0)$ (A.8.4).

In order to define $C^{(2)}$, we should first define $A^{(2)}$. For simplicity, we define two states α and β for each bond here and we have 2^2x 2^2 dimensional matrixes and every state has its own transition rate.

αα →αα r_{11}	αβ → αα r_{21}	βα → αα r_{31}	ββ → αα r_{41}
αα→ αβ r_{12}	αβ → αβ r_{22}	βα → αβ r_{32}	ββ → αβ r_{42}
αα→ βα r_{13}	αβ → βα r_{23}	βα → βα r_{33}	ββ → βα r_{43}
αα→ ββ r_{14}	αβ → ββ r_{24}	βα → ββ r_{34}	ββ → ββ r_{44}

Transition rates matrices are defined as follows;

$$A = from\begin{bmatrix} -r_{21}-r_{31}-r_{41} & r_{12} & r_{13} & r_{14} \\ r_{21} & -r_{12}-r_{32}-r_{42} & r_{23} & r_{24} \\ r_{31} & r_{32} & -r_{13}-r_{23}-r_{43} & r_{34} \\ r_{41} & r_{42} & r_{43} & -r_{14}-r_{24}-r_{34} \end{bmatrix}$$

$$to$$

Finally, we have the joint probability matrix for bond *i* as a function of time and defined as $p_{\zeta\eta(t),\zeta\eta(o);i}$.In addition, we define the probability for bond i-1 in the state ζ at time 0 and time t as $p_{\zeta(t),\zeta(o);i-1}$.The partition function for transition probabilities is denoted by $\mathrm{Z}(\tau)$ for a given time interval, τ . Conventionally, the partition function is the serial multiplication of stochastic weight matrices as follows;

$$Z(\tau) = V_1(\tau)V_2(\tau)....V_{n-1}(\tau)V_n(\tau) \qquad \text{(A.8.5)}$$

The elements of the stochastic weight matrix are denoted as *v* and we define it as

$$v = \frac{p_{\zeta\eta(t),\zeta\eta(o);i}}{p_{\zeta(t),\zeta(o);i-1}} \qquad \text{(A.8.6)}$$

Moreover, the stochastic weight matrix should be rearranged in order to ensure that serial multiplication of the matrices respects the chain connectivity. *V$_i$(τ)* is divided into submatrices, each represents the stochastic weights for the transitions to a given final state.

$$V_i(\tau) = \left[\begin{array}{cc:cc} v_i(\alpha\alpha;\alpha\alpha) & v_i(\alpha\alpha;\alpha\beta) & v_i(\alpha\beta;\alpha\alpha) & v_i(\alpha\beta;\alpha\beta) \\ v_i(\alpha\alpha;\beta\alpha) & v_i(\alpha\alpha;\beta\beta) & v_i(\alpha\beta;\beta\alpha) & v_i(\alpha\beta;\beta\beta) \\ \hline v_i(\beta\alpha;\alpha\alpha) & v_i(\beta\alpha;\alpha\beta) & v_i(\beta\beta;\alpha\alpha) & v_i(\beta\beta;\alpha\beta) \\ v_i(\beta\alpha;\beta\alpha) & v_i(\beta\alpha;\beta\beta) & v_i(\beta\beta;\beta\alpha) & v_i(\beta\beta;\beta\beta) \end{array}\right] \qquad \text{(A.8.7)}$$

The probability of the transition from one state to another can be determined from the joint probability calculations;

$$p_{\zeta\eta(t),\zeta\eta(0),i} = Z(\tau)V_1(\tau)V_2(\tau).....V_{k-1}V_k^{'}V_{k+1}.....V_{n-1}(\tau)V_n(\tau) \qquad \text{(A.8.8)}$$

Where $V^{'}$ is the stochastic weight matrix obtained by equating the entries zero except that *v$_i$(ζη(t), ζη(0)).*

Orientational Autocorrelation Function

Local orientational motions depend in fact on the transition of several consecutive bonds. A quantitative measure of such motions would be the orientational autocorrelation function (OACF) related with a vectoral quantity m rigidly affixed to the chain. The function is defined as $< m(0) \cdot m(\tau) >$. (A.8.8)

The OACF which depends on the rotations of the pair bonds (i-1,i) is

$$< m(0)\cdot m(\tau) >= m^{0T} < T_i^T(0)T_{i-1}^T(0)T_{i-1}(\tau)T_i\ (\tau) > m^0 \qquad \text{(A.8.9)}$$

where T_i is the transformation matrix that express m in the ith bond-based local frame. The average over all configuration transitions can be calculated by using the following mathematical method that is formulated before.

We define a pseudo diagonal matrix

$$\|S_i\| = \begin{bmatrix} T_i^T(\alpha^0)T_{i-1}^T(\alpha^0)T_{i-1}(\alpha\ )T_i^T(\alpha) & & & \\ & \cdots\cdots\cdots\cdots & & \\ & & \cdots\cdots\cdot & \\ & & & T_i^T(\beta^0)T_{i-1}^T(\beta^0)T_{i-1}(\beta\ )T_i^T(\beta) \end{bmatrix}$$

Then

$$< T_i^T(0)T_i(\tau) >= Z^{-1}(J^* \otimes I_n)\left[\prod_{j=2}^{i-1} V_j(\tau)\otimes I_n\right]\left[V_j(\tau)\otimes I_n\|S_i\|\right]\left[\prod_{k=i+1}^{N-1} V_k(\tau)\otimes I_n\right](J\ \ \otimes I_n)$$

(A.8.10)

BIBLIOGRAPHY

[1]. Horton, H.R., L.A. Moran, R.S. Ochs, and J.D. Rawn, *Principles of Biochemistry*. 2002: Prentice-Hall, Inc.

[2]. Ramakrishnan, C. and G.N. Ramachandran, Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. *Biophys J*, 1965. 5(6): p. 909-33.

[3]. Staniforth, R.A., et al., The energetics and cooperativity of protein folding: a simple experimental analysis based upon the solvation of internal residues. *Biochemistry*, 1993. 32(15): p. 3842-51.

[4]. Ghelis, C. and J. Yon, *Protein Folding*. 1982, New York: Academic Press.

[5]. Bryngelson, J.D., J.N. Onuchic, N.D. Socci, and P.G. Wolynes, Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins*, 1995. 21(3): p. 167-95.

[6]. Dill, K.A. and H.S. Chan, From Levinthal to pathways to funnels. *Nat Struct Biol*, 1997. 4(1).

[7]. Baldwin, L.R. and G.D. Rose, Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends in Biochemical Sciences*, 1999. 24(2): p. 77-83.

[8]. Billing, G.D., Advanced molecular dynamics and chemical kinetics. 1997.

[9]. Rapaport, D.C., The Art of Molecular Dynamics Simulation. 1997.

[10]. Ramachandran, G.N., C. Ramakrishnan, and V. Sasisekharan, Stereochemistry of polypeptide chain conformations. *J Mol Biol*, 1963. 7: p. 95-99.

[11]. Penkett, C.J., et al., NMR analysis of main-chain conformational preferences in an unfolded fibronectin-binding protein. *J Mol Biol*, 1997. 274(2): p. 152-9.

[12]. Smith, L.J., K.A. Bolin, H. Schwalbe, M.W. MacArthur, J.M. Thornton, and C.M. Dobson, Analysis of main chain torsion angles in proteins: prediction of NMR coupling constants for native and random coil conformations. *J Mol Biol*, 1996. 255(3): p. 494-506.

[13]. Bahar, I., M. Kaplan, and R.L. Jernigan, Short-Range Conformational Energies, Secondary Structure Propensities, and Recognition of Correct Sequence-Structure Matches. *PROTEINS: Structure, Function and Bioinformatics*, 1997. 29: p. 292-308.

[14]. Karplus, P.A., Experimentally observed conformation-dependent geometry and hidden strain in proteins. *Protein Sci*, 1996. 5(7): p. 1406-20.

[15]. Jha, A.K., A. Colubri, M.H. Zaman, S. Koide, T.R. Sosnick, and K.F. Freed, Helix, Sheet, and Polyproline II Frequencies and Strong Nearest Neighbor Effects in a Restricted Coil Library. *Biochemistry*, 2005. 44(28): p. 9691-702.

[16]. Avbelj, F. and R.L. Baldwin, Origin of the neighboring residue effect on peptide backbone conformation. *Proc Natl Acad Sci U S A*, 2004. 101(30): p. 10967-72.

[17]. Berman, H.M., et al., The Protein Data Bank. *Nucleic Acids Res*, 2000. 28(1): p. 235-42.

[18]. Keskin, O., D. Yuret, A. Gursoy, M. Turkay, and B. Erman, Relationships Between Aminoacid Sequence and Backbone Torsion Angle Preferencse. *PROTEINS: Structure, Function and Bioinformatics*, 2004. 55(4): p. 992-998.

[19]. Jackson, S.E. and A.R. Fersht, Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry*, 1991. 30(43): p. 10428-35.

[20]. Jackson, S.E., M. Moracci, N. elMasry, C.M. Johnson, and A.R. Fersht, Effect of cavity-creating mutations in the hydrophobic core of chymotrypsin inhibitor 2. *Biochemistry*, 1993. 32(42): p. 11259-69.

[21]. Jackson, S.E. and A.R. Fersht, Folding of Chymotrypsin Inhibitor 2. 2. Influence of Proline Isomerization on theFolding Kinetics and Thermodynamic Characterization of the transition State of Folding. *Biochemistry*, 1991. 30: p. 10436-10443.

[22]. Kaya, H. and H.S. Chan, Polymer principles of protein calorimetric two-state cooperativity. *PROTEINS: Structure, Function and Bioinformatics*, 2000. 40: p. 637-661.

[23]. Micheletti, C., J.R. Banavar, and A. Maritan, Conformations of proteins in equilibrium. *Phys Rev Lett*, 2001. 87(8): p. 088102.

[24]. Flory, P.J., Foundations of Rotational Isomeric StateTheory and General Methods for Generating Configurational Averages. *Macromolecules*, 1974. 7: p. 381.

[25]. Day, R. and V. Daggett, Sensitivity of the folding/unfolding transition state ensemble of chymotrypsin inhibitor 2 to changes in temperature and solvent. *Protein Sci*, 2005. 14(5): p. 1242-52.

[26]. Lazaridis, T. and M. Karplus, Heat Capacity and compactness of denaturated proteins. *Biophys Chemistry*, 1999. 78: p. 207-217.

[27]. Hao, M.H. and H.A. Scheraga, On foldable protein-like models; statistical-mechanical study with Monte Carlo Simulations. *Pysics A*, 1997. 244: p. 124-146.

[28]. Eaton, W.A., V. Munoz, P.A. Thompson, C.K. Chan, and J. Hofrichter, Submillisecond kinetics of protein folding. *Curr Opin Struct Biol*, 1997. 7(1): p. 10-4.

[29]. Fersht, A.R., Nucleation mechanisms in protein folding. *Curr Opin Struct Biol*, 1997. 7(1): p. 3-9.

[30]. Forcellino, F. and P. Derreumaux, Computer simulations aimed at structure prediction of supersecondary motifs in proteins. *Proteins*, 2001. 45(2): p. 159-66.

[31]. Elizier, D., J. Yao, H.J. Dyson, and P.E. Wright, Structural and dynamic characterization of partially folded states of apomyoglobin and implications for protein folding. *Nat. Structural Biology*, 1998. 5(2).

[32]. Markwick, P.R., R. Sprangers, and M. Sattler, Local structure and anisotropic backbone dynamics from cross-correlated NMR relaxation in proteins. *Angew Chem Int Ed Engl*, 2005. 44(21): p. 3232-7.

[33]. Skolnick, J. and A. Kolinski, Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure and dynamics. *J Mol Biol*, 1992. 223(2): p. 583.

[34]. Ding, F., R.K. Jha, and N.V. Dokholyan, Scaling behavior and structure of denatured proteins. *Structure (Camb)*, 2005. 13(7): p. 1047-54.

[35]. Krieger, F., B. Fierz, O. Bieri, M. Drewello, and T. Kiefhaber, Dynamics of unfolded polypeptide chains as model for the earliest steps in protein folding. *J Mol Biol*, 2003. 332(1): p. 265-74.

[36]. Bahar, I. and B. Erman, Investigation of Local Motions in Polymers by the Dynamic Rotational Isomeric State Model. *Macromolecules*, 1987. 20(6): p. 1368-1376.

[37]. Bahar, I. and B. Erman, Comparison of Dynamics Rotational Isomeric State Results with Previous Expressions for Local Chain Motions. *Macromolecules*, 1989. 22(1): p. 431-437.

[38]. Bahar, I., B. Erman, and L. Monnerie, Application of the Dynamics Rotational Isomeric States Model to Poly(ethylene oxide) and Comparison with Nuclear Magnetic Relaxation Data. *Macromolecules*, 1989. 22(5).

[39]. Bahar, I. and W.L. Mattice, Efficient Calculation of the Intramolecular Contributiın to Orientational Autocorrelation Functions Using Dynamic Rotational Isomeric State Theory. *Macromolecules*, 1990. 23(10).

[40]. Humphrey, W., A. Dalke, and K. Schulten, {VMD} -- {V}isual {M}olecular {D}ynamics. 1996. 14: p. 33-38.

[41]. Pearlmann, D.A., et al., AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. *Comp. Physc. Commun.*, 1995. 91: p. 1-41.

[42]. Brooks, B.R., B. R.E., B.D. Olafson, D.J. States, S. Swaminathan, and M. Karplus, CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistery*, 1983. 4(2): p. 187-217.

[43]. Brünger, A.T., *X-PLOR, Version 3.1, A System for X-ray Crystallography and NMR.* 1992: TheHoward Hughes Medical Institute and Department of Molecular Biophysics and Biochemistry, Yale University.

[44]. Laxmikant, K., et al., NAMD2: Greater scalability for parallel molecular dynamics. *Journal of Computational Physics*, 1999. 151: p. 283-312.

[45]. Fanelli, F., C. Menziani, A. Scheer, S. Cotecchia, and P.G. De Benedetti, Theoretical study of the electrostatically driven step of receptor-G protein recognition. *Proteins*, 1999. 37(2): p. 145-56.

[46]. Allen, M.P. and D.J. Tildesley, *Computer Simulation of Liquids*. 1987, New York: Oxford University Press.

[47]. Flory, P.J., *Statistical Mechanics of Chain Molecules*. 1969, New York: Wiley.

[48]. Wall, M. *GAlib:A C++Library of Genetic Algorithm Components*. 1996 [cited; Available from: http://lancet.mit.edu/ga/.

[49]. Ramakrishnan, C. and D.V. Nataraj, Energy minimization studies on alpha-turns. *J Pept Sci*, 1998. 4(4): p. 239-52.

[50]. Serrano, L., Comparison between the phi distribution of the amino acids in the protein database and NMR data indicates that amino acids have various phi propensities in the random coil conformation. *J Mol Biol*, 1995. 254(2): p. 322-33.

[51]. Callen, H.B., Thermodynamics and an Introduction to Thermostatistics. 1960.

[52]. Itzhaki, L.S., D.E. Otzen, and A.R. Fersht, The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J Mol Biol*, 1995. 254(2): p. 260-88.

[53]. Otzen, D.E., L.S. Itzhaki, N.F. elMasry, S.E. Jackson, and A.R. Fersht, Structure of the transition state for the folding/unfolding of the barley chymotrypsin inhibitor 2 and its implications for mechanisms of protein folding. *Proc Natl Acad Sci U S A*, 1994. 91(22): p. 10422-5.

[54]. Jackson, S.E., N. elMasry, and A.R. Fersht, Structure of the hydrophobic core in the transition state for folding of chymotrypsin inhibitor 2: a critical test of the protein engineering method of analysis. *Biochemistry*, 1993. 32(42): p. 11270-8.

[55]. Makhatadze, G.I. and P.L. Privalov, Energetics of protein structure. *Adv Protein Chem*, 1995. 47: p. 307-425.

[56]. Makhatadze, G.I. and P.L. Privalov, On the entropy of proetin folding. *The Protein Society*, 1996. 5: p. 507-510.

[57]. Lee, K.H., D. Xie, E. Freire, and L.M. Amzel, Estimation of changes in side chain configurational entropy in binding and folding: general methods and application to helix formation. *Proteins*, 1994. 20(1): p. 68-84.

[58]. Bryngelson, J.D. and P.G. Wolynes, Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci U S A*, 1987. 84(21): p. 7524-8.

[59]. Nemethy, G., S. Leach, and H.A. Scheraga, The influence of amino acid side chains on the free energy of helix-coil transitions. *J Phys Chem*, 1966. 70(998-1004).

[60]. Miller, W.G., D.A. Brant, and P.J. Flory, Random Coil Configuration of Polypeptide Copolymers. *Journal of Molecular Biology*, 1967. 23: p. 67-80.

[61]. Mattice, W.L. and U.W. Suter, *Conformational Theory of Large Molecules. The Rotational Isomeric State Model in Macromolecular Systems*. 1994, New York: John Wiley & Sons.

[62]. Chan, H.S. and K.A. Dill, Polymer principles in protein structure and stabilitiy. *Annu. Rev. Biophys. Biophys. Chem*, 1991. 20: p. 447-490.

VITA

Ayşe Meriç Ovacık was born in Izmir, on April 26, 1980. She received her B.Sc. degree in Chemical Engineering from Middle East Technical University, Ankara, in 2003. From September to present she has been worked as teaching and research assistant in Koc University, Istanbul. She assisted physics 102&102, chemistry 102 and biology 200 courses and studied to develop “STATISTICAL MECHANICS AND LOCAL DYNAMICS OF DENATURATED PROTEINS” project.